

# Topic 9: 机器学习原理及应用

## Machine Learning Fundamentals and Applications

刘跃文 博士 Dr. LIU, Yuewen

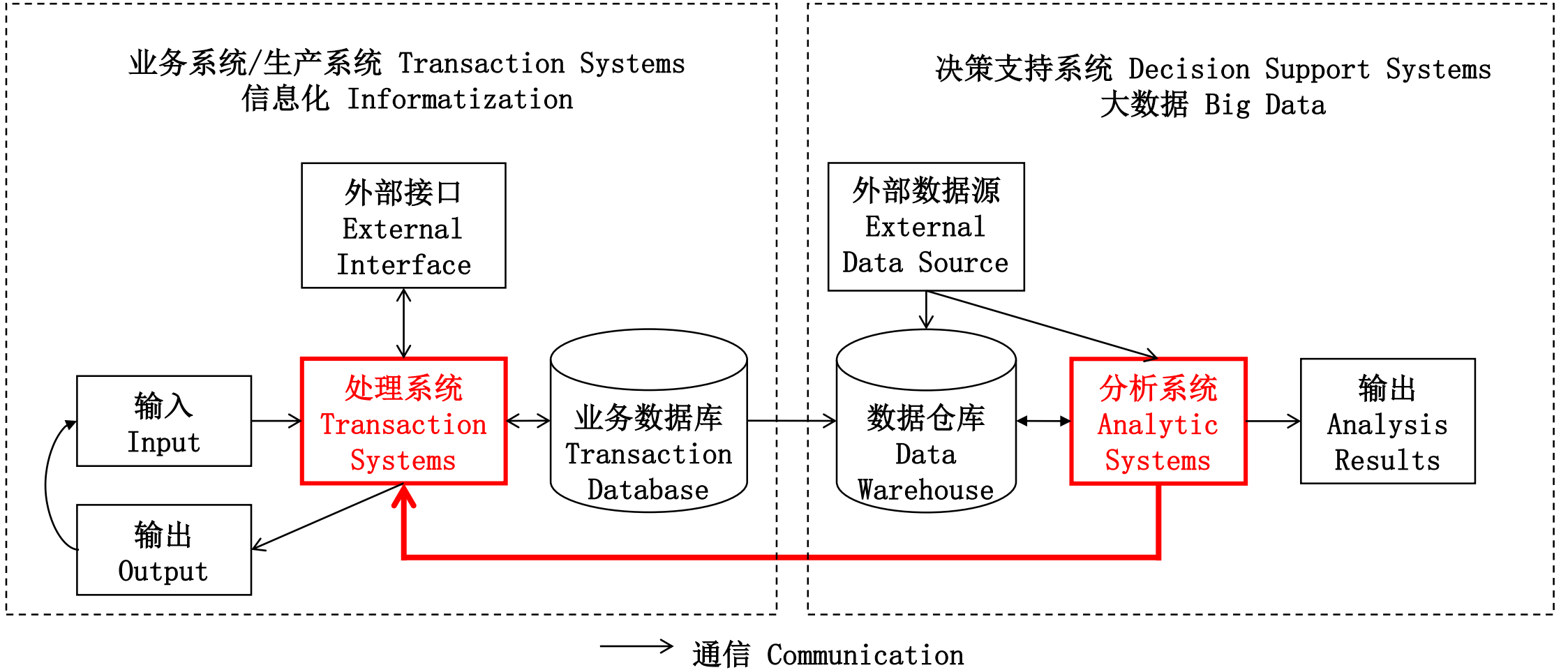
教授、博士生导师 Professor

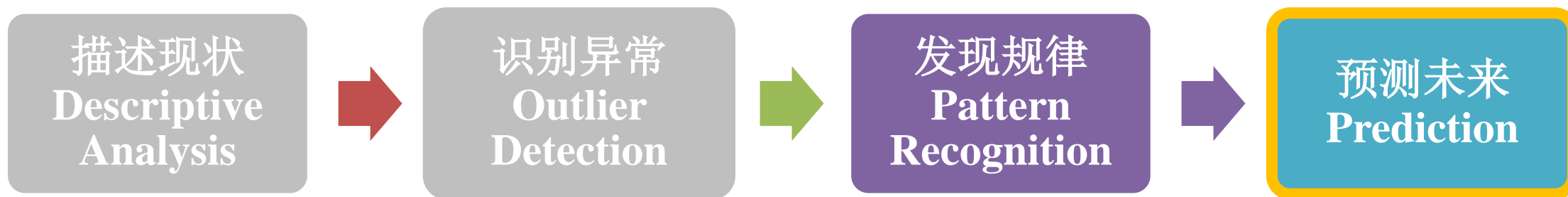
[liuyuewen@xjtu.edu.cn](mailto:liuyuewen@xjtu.edu.cn)

西安交通大学管理学院

School of Management, Xi'an Jiaotong University

V2.0, 2023-Oct





**数据挖掘：知识发现**  
**Data Mining: Knowledge Discovery**

# 提纲Outline

---

1. 机器学习原理Machine Learning Fundamentals
2. 特征表与数据预处理Feature Table and Data Preprocessing
3. 机器学习模型的分类与应用场景Classification and Application of Machine Learning
4. 机器学习模型的评估与弱点Evaluation of Machine Learning Models
5. 机器学习与商业模式Machine Learning and Business Models

# 1. 机器学习原理

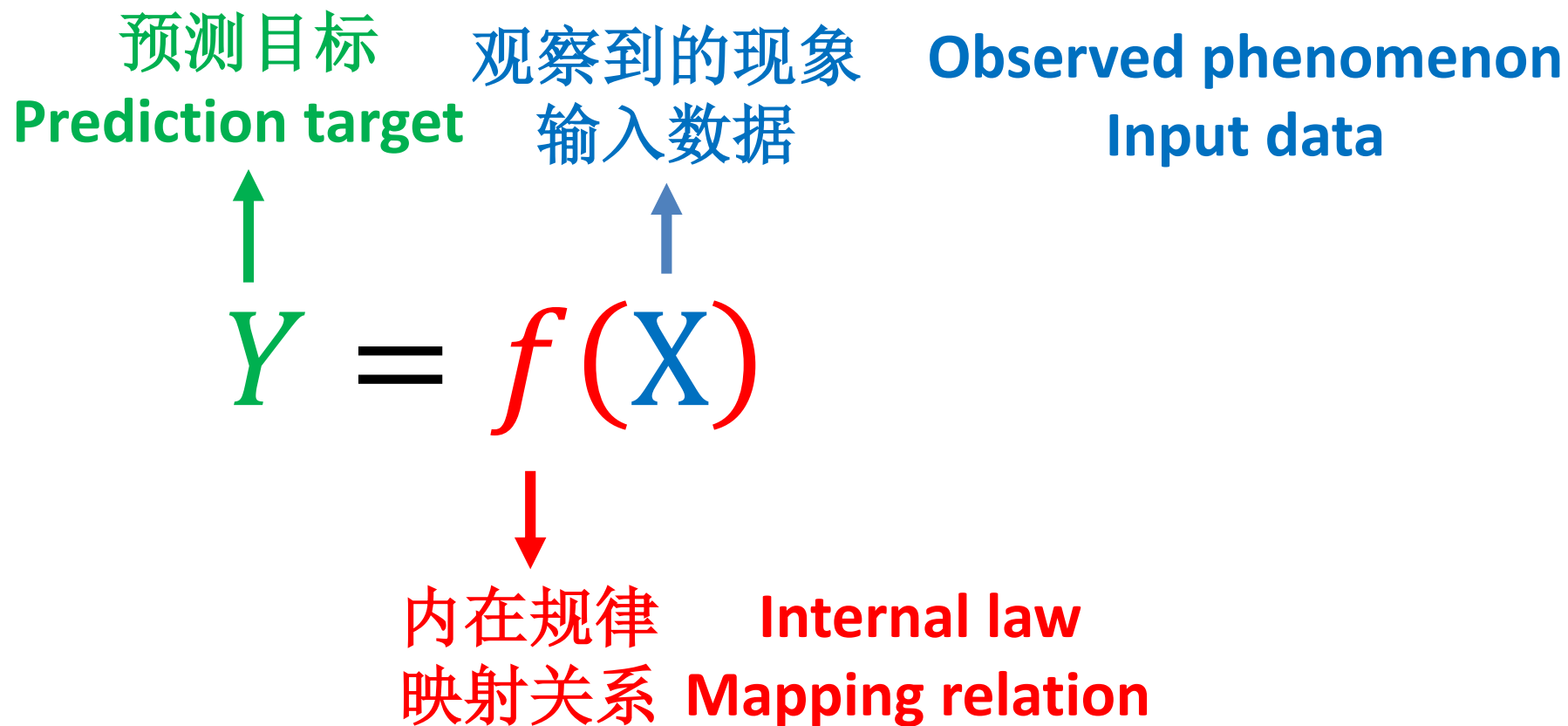
## Machine Learning Fundamentals



# 1. The essence is to find functions

- Mapping/calculation mechanism from Input to Output.

ML Models	Input X	$f()$	Output Y
Customer Churn Prediction	Historical customer data	→	Customer churn/non-churn
Customer Value Prediction	Customer consumption data	→	Customer Value
Image recognition	Animal images	→	Image labeling: Cat/Dog
Face recognition	Face images	→	identity
Machine translation	Chinese	→	English
Autopilot	Radar data	→	Vehicle position

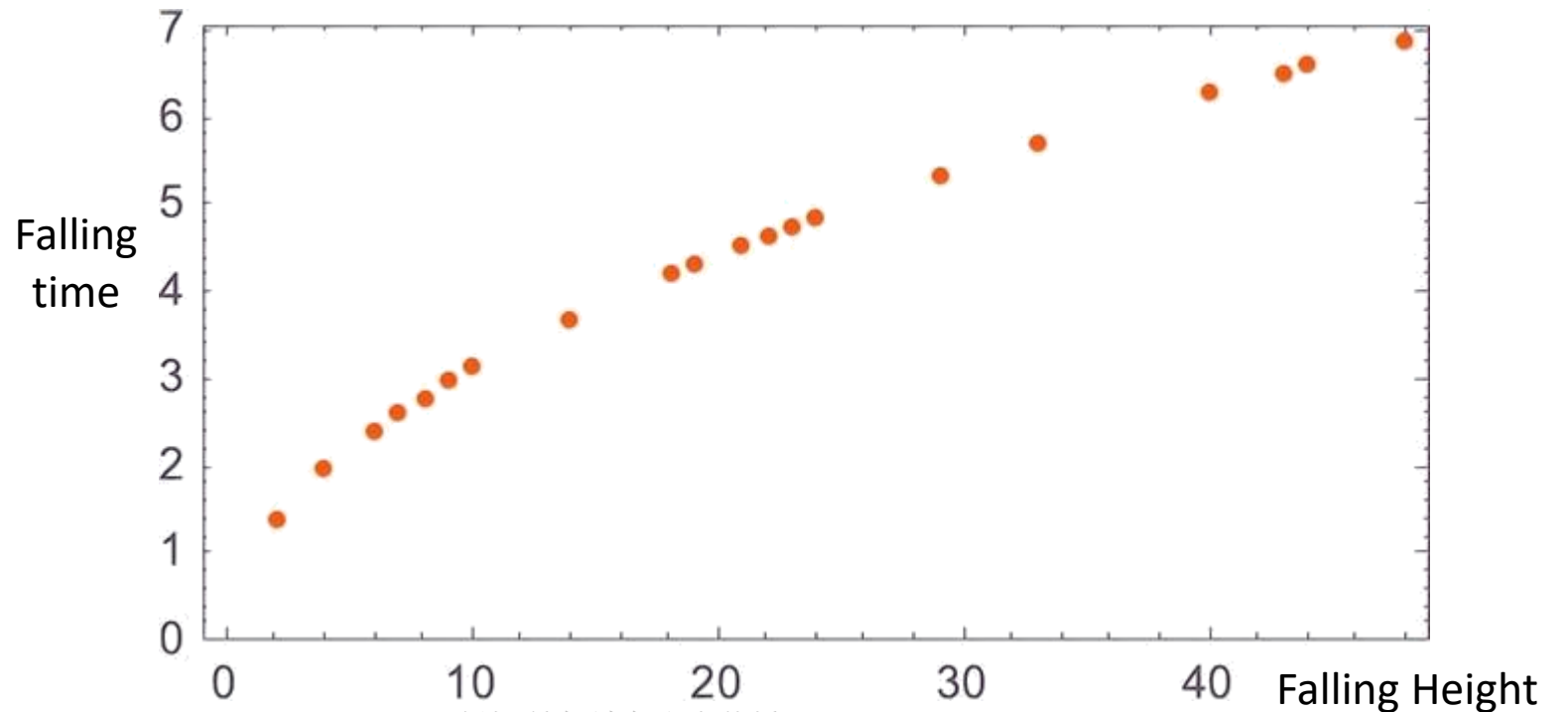


## 2. Predict the landing time of the iron ball

---



Suppose you're in the late 16th century and you want to know how long it takes for a cannonball falling from each floor of the Leaning Tower of Pisa to hit the ground. You can measure it in each case and make a resulting chart.





Prediction target 预测目标  
Falling time 下落时间



$Y$

$=$

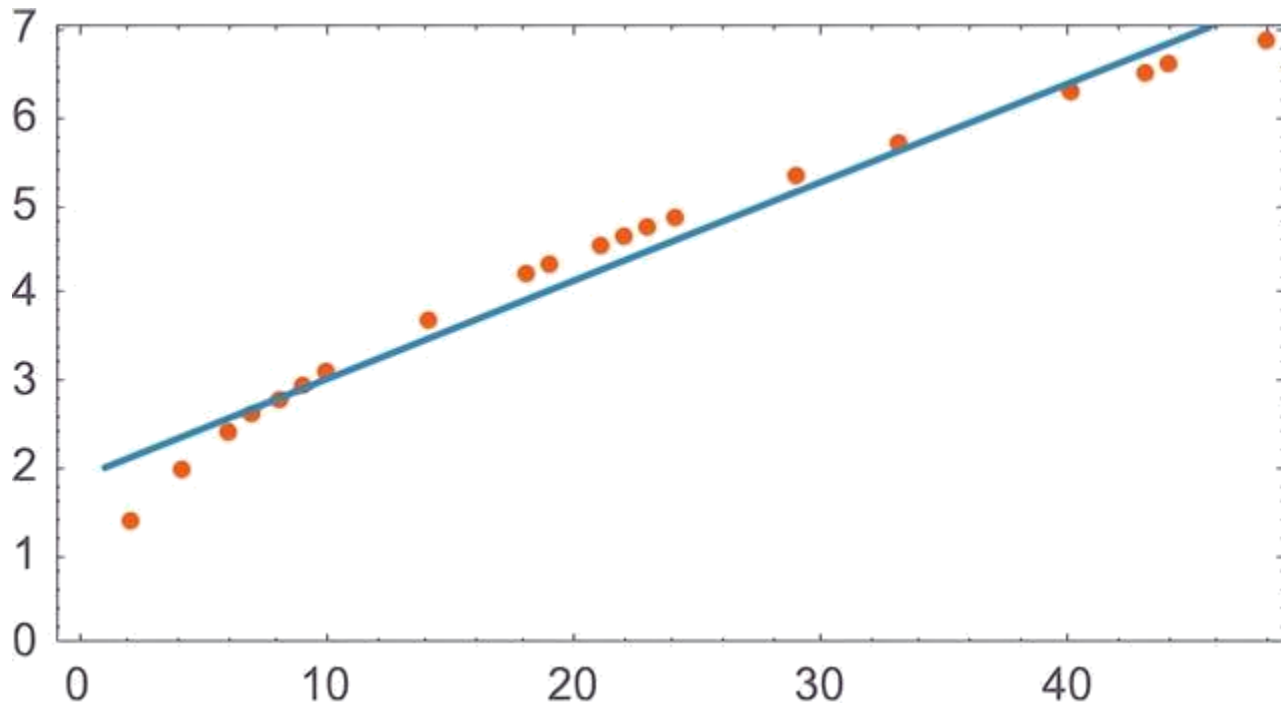
$f(X)$

输入数据 Input data  
下落高度 Falling height



**Step 1:** Find a suitable functional form.

**Step 2:** Estimate (learn) the magnitude of the parameters based on the data.



**Prediction target**  $Y$  – *The falling time of the ball*

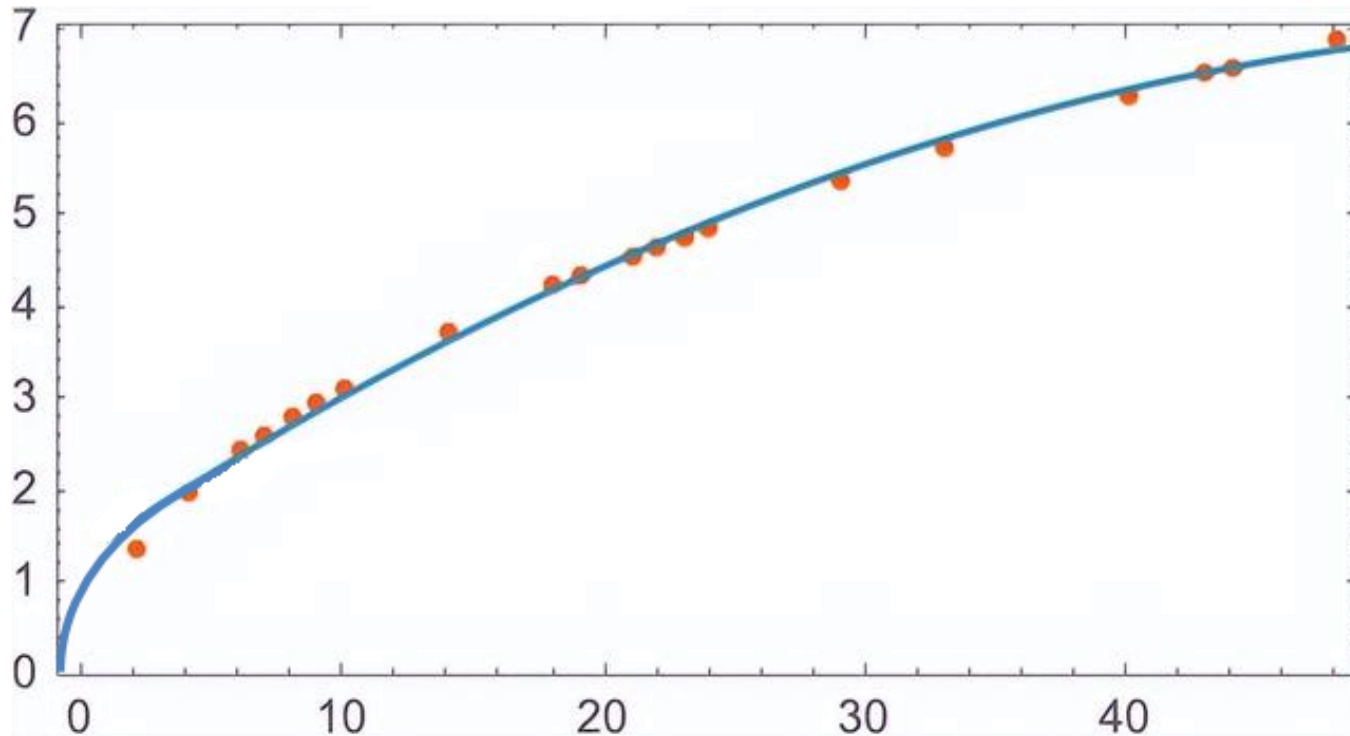
**Input Variable**  $X$  – *The Falling height of the ball*

**Function**  $Y = aX + b$

**Fitting function**  $Y = 0.1X + 2$

**Learning parameters**  $a = 0.1, b = 2$

choose the **"right"** form of the function;  
understand the **theory/rationale** behind functional forms.



**Function Form**

$$Y = a\sqrt{X} + b$$

**Fitting**

$$Y = 0.45\sqrt{X}$$

**Learning  
Parameters**

$$a = 0.45, b = 0$$

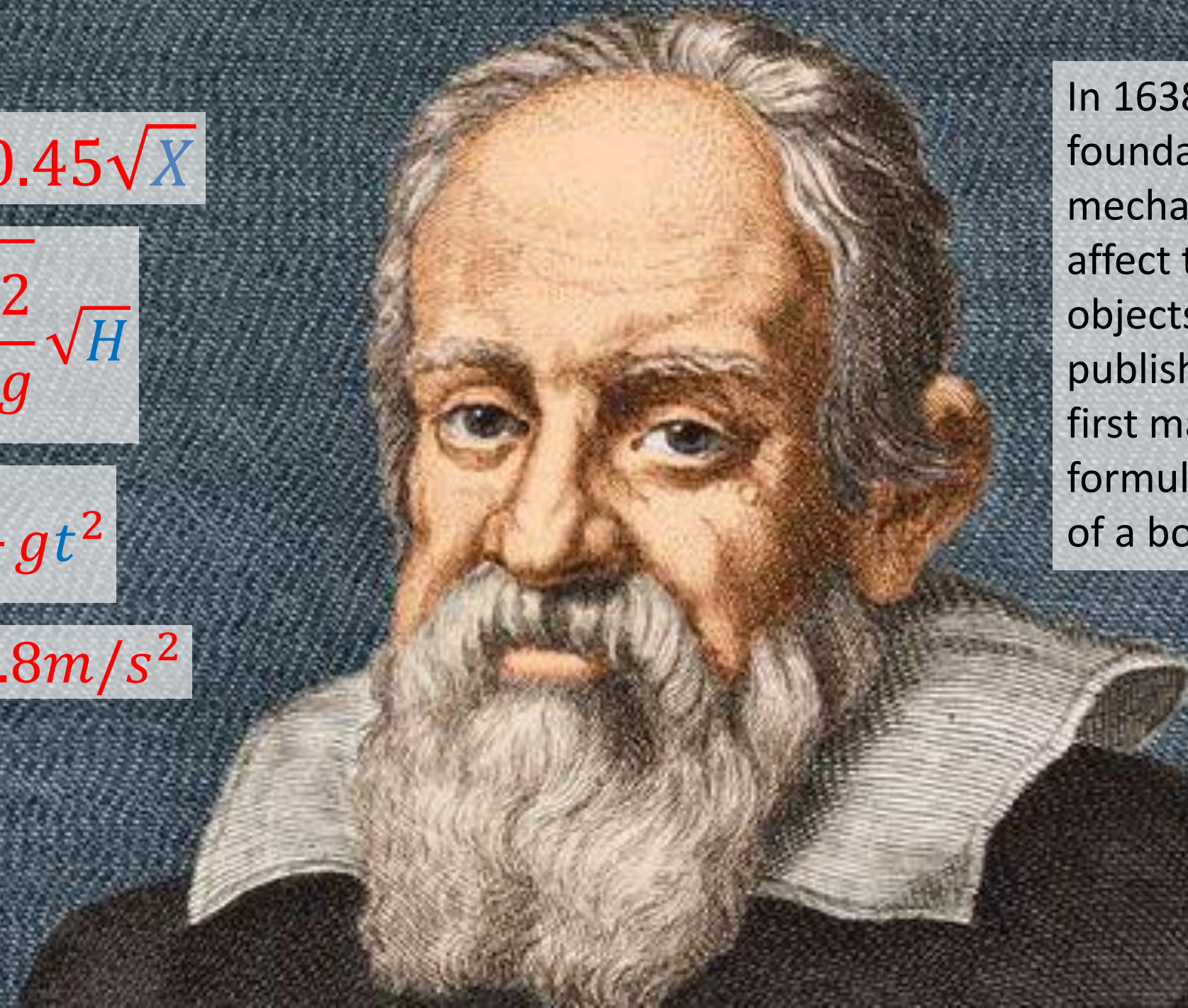


$$Y = 0.45\sqrt{X}$$

$$t = \sqrt{\frac{2}{g}}\sqrt{H}$$

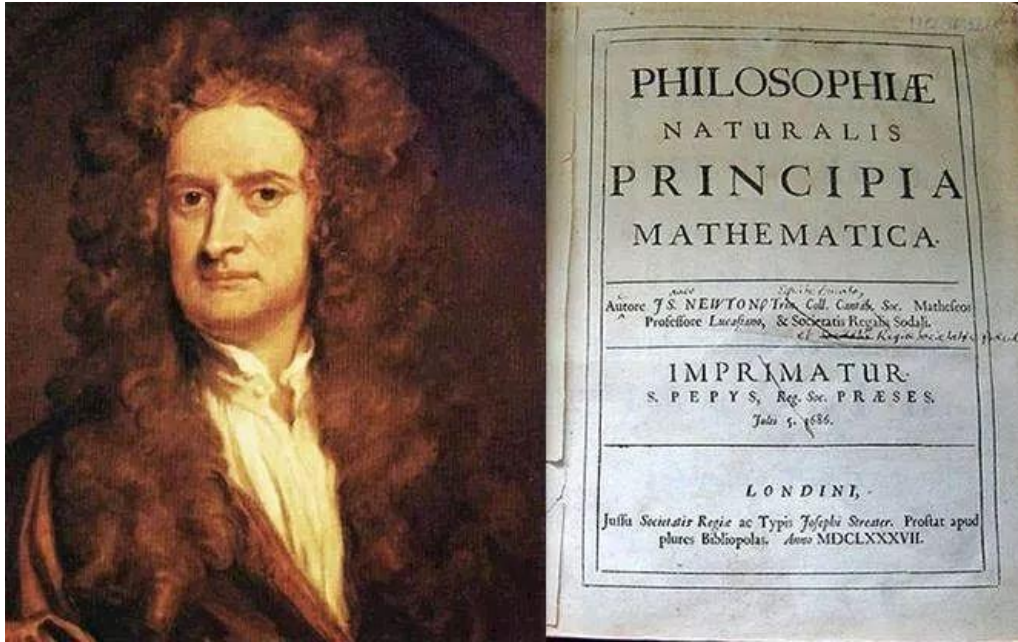
$$H = \frac{1}{2}gt^2$$

$$g = 9.8\text{m/s}^2$$



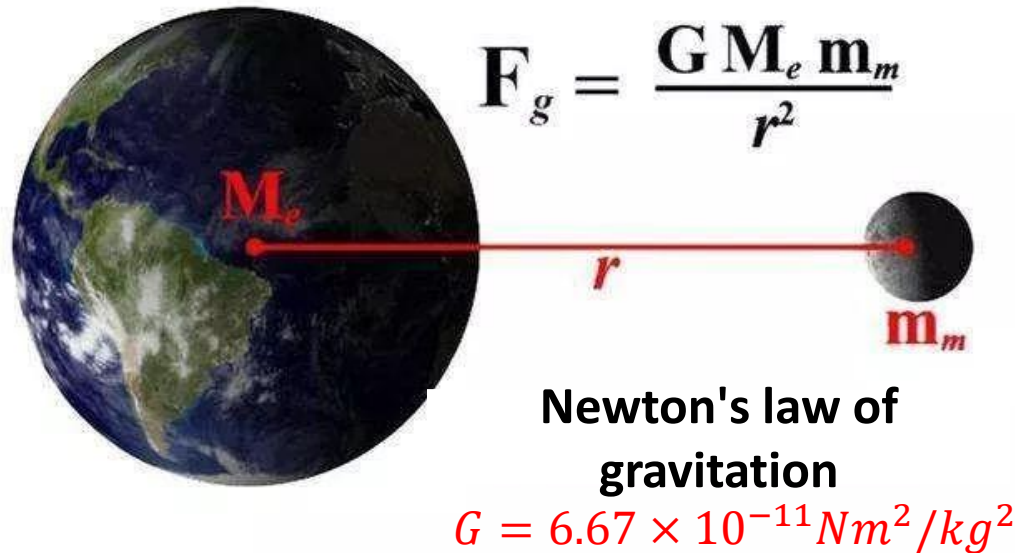
In 1638, Galileo laid the foundations of classical mechanics -- how forces affect the motion of objects -- in a paper he published. He gave the first mathematical formula for the motion of a body in free fall.





In 1687 (the 26th year of Kangxi in Qing Dynasty), the British physicist Isaac. Newton published his magnum opus, Principia Mathematica, which heralded the coming of the scientific age.

In the third volume of Principia Mathematica, Newton wrote: "Finally, if it is generally shown by experiment and astronomical observation that all the celestial bodies around the Earth are attracted by the gravity of the Earth, and that their gravity is in proportion to the amount of matter they contain, then the moon is likewise attracted by the gravity of the Earth in proportion to the amount of matter they contain." On the other hand, it shows that our oceans are drawn to the moon's gravity; And all the planets are attracted to each other by gravity, and comets are also attracted to the sun's gravity. Because of this rule, we must generally admit that all bodies, whatever they may be, are endowed with the principle of mutual gravitation. For the argument for the universal gravitation of all bodies derived from this representation..."



### 3. Anticipate/alert customer churn

---

- Customer churn prediction
- The development of a new customer requires a certain cost, once the customer loss, will cause losses to the business, so the prediction of customer loss dose matters.
- The role of predicting customer churn:
  - Predict which customers are likely to churn and use touch strategies to retain customers before they churn;
  - Analyze the reasons for customer churn and look for leading indicators to increase retention and improve the product.
  - For the lost customers, change the product operation strategy to pull back customers and promote reflux.

<https://www.jianshu.com/p/143782bc15e4>

<https://zhuanlan.zhihu.com/p/40197660>

<https://zhuanlan.zhihu.com/p/68397317>



As a bank/carrier account manager, you aim to identify potential customer attrition within the next few months to proactively retain them. You can access extensive customer data from the company's information system for this purpose.

	A	B	C	D	E	U
1	customerID	gender	SeniorCitizen	Partner	Dependents	Churn
2	7590-VHVEG	Female	0	Yes	No	No
3	5575-GNVDE	Male	0	No	No	No
4	3668-QPYBK	Male	0	No	No	Yes
5	7795-CFOCW	Male	0	No	No	No
6	9237-HQITU	Female	0	No	No	Yes
7	9305-CDSKC	Female	0	No	No	Yes
8	1452-KIOVK	Male	0	No	Yes	No
9	6713-OKOMC	Female	0	No	No	No
10	7892-POOKP	Female	0	Yes	No	Yes
11	6388-TABGU	Male	0	No	Yes	No
12	9763-GRSKD	Male	0	Yes	Yes	No
13	7469-LKBCI	Male	0	No	No	No
14	8091-TTVAX	Male	0	Yes	No	No
15	0280-XJGEX	Male	0	No	No	Yes

Predict whether  
the target is lost

预测目标  
是否流失

$Y = f($

Variables/attributes/features

- gender 性别 (male/female)
- SeniorCitizen (老年人与否1/0)
- Partner (有无合作伙伴)
- Dependents (有无家属)
- tenure (用户入网月数/留存月数)
- PhoneService (用户是否有电话服务)
- MultipleLines (用户是否有多线)
- InternetService (互联网服务提供商: DSL, Fiber optic, No)
- OnlineSecurity (在线安全Yes, No, No internet service)
- OnlineBackup(在线备份Yes, No, No internet service)
- DeviceProtection (设备检测Yes, No, No internet service)
- TechSupport (技术支持Yes, No, No internet service)
- StreamingTV (流媒体电视Yes, No, No internet service)
- StreamingMovie (流媒体电影Yes, No, No internet service)
- Contract (客户的合同期限: Month-to-month, One year, Two year)
- PaperlessBilling(无纸化账单: Yes, No)
- PaymentMethod (支付方式: Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
- MonthlyCharges (月费用)
- TotalCharges (总费用)

$X$

)



# Simple business experience summary and expression

---

- Simple business experience summary: What are the characteristics of lost customers?
  - For example: male, middle-aged, 0-60 calls per month, 0-300 minutes per month.
- Simple business experience expression/application:
  - **Conditional Filtering:** Identify potential churn customers by selecting males aged 30-50 with monthly call frequencies between 0-60 times and call durations of 0-300 minutes for proactive customer care.
  - **Scoring Alert:** Assign scores based on criteria such as **gender** [20 points for males, 0 points for females], **age** [20 points for 30-50 years, 10 points for 20-30 years, 10 points for 50-60 years, 0 points for other], **monthly call frequency** [20 points for 0-60 times, 10 points for 60-150 times, 0 points for other], **monthly call duration** [20 points for 0-300 minutes, 10 points for 300-500 minutes, 0 points for other]... A higher cumulative score indicates higher risk, and if it exceeds 100 points, the customer is flagged as a potential churn candidate for alert.

# Conditional filtering method

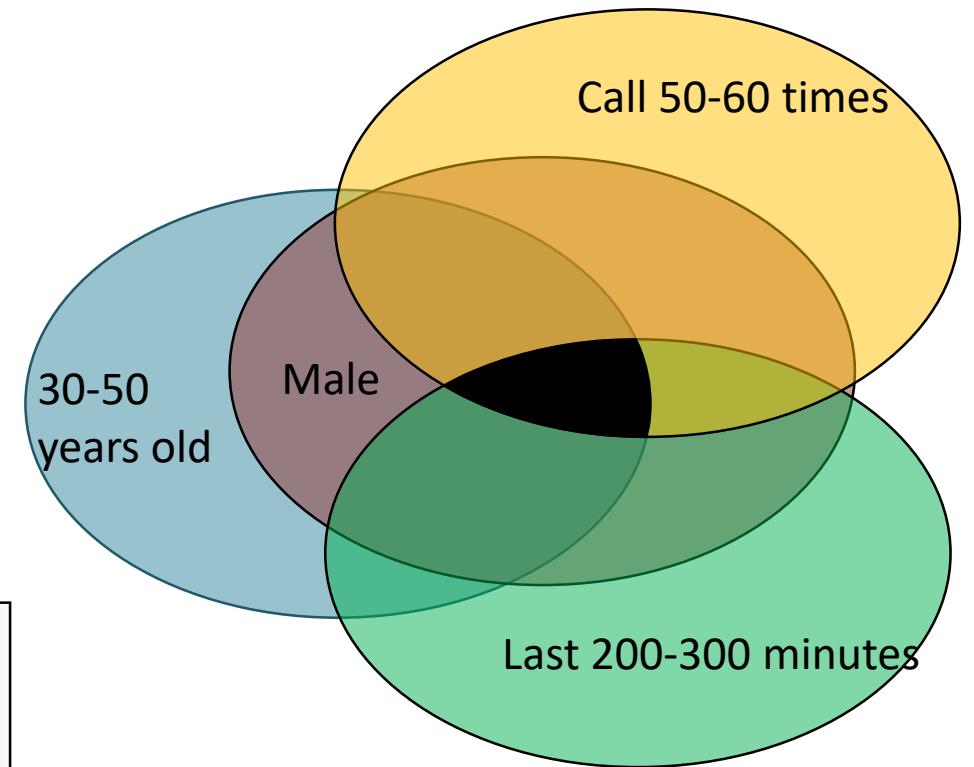
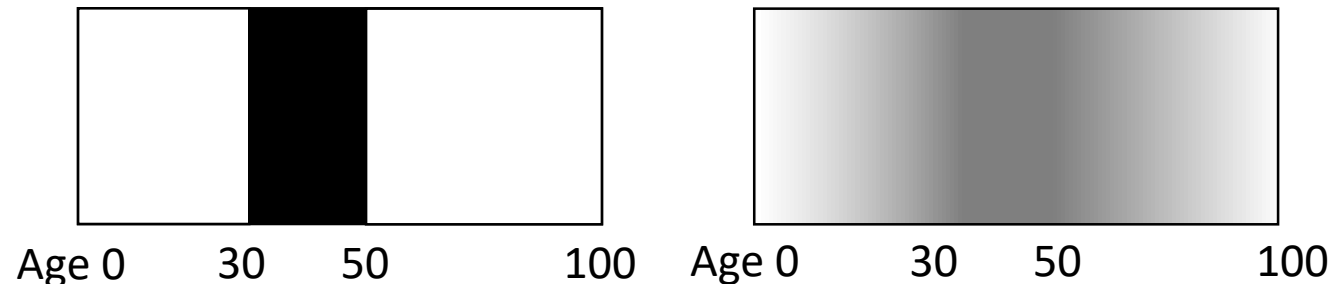
---

$$Y = \begin{cases} 1 & \text{if } gender = 1 \text{ and } age \geq 30 \\ & \text{and } age \leq 50 \text{ and } calls \leq 60 \\ & \text{and } callLength \leq 300; \\ 0 & \text{otherwise.} \end{cases}$$

Issue: Parameters and thresholds are arbitrarily determined; controlling the score is challenging.

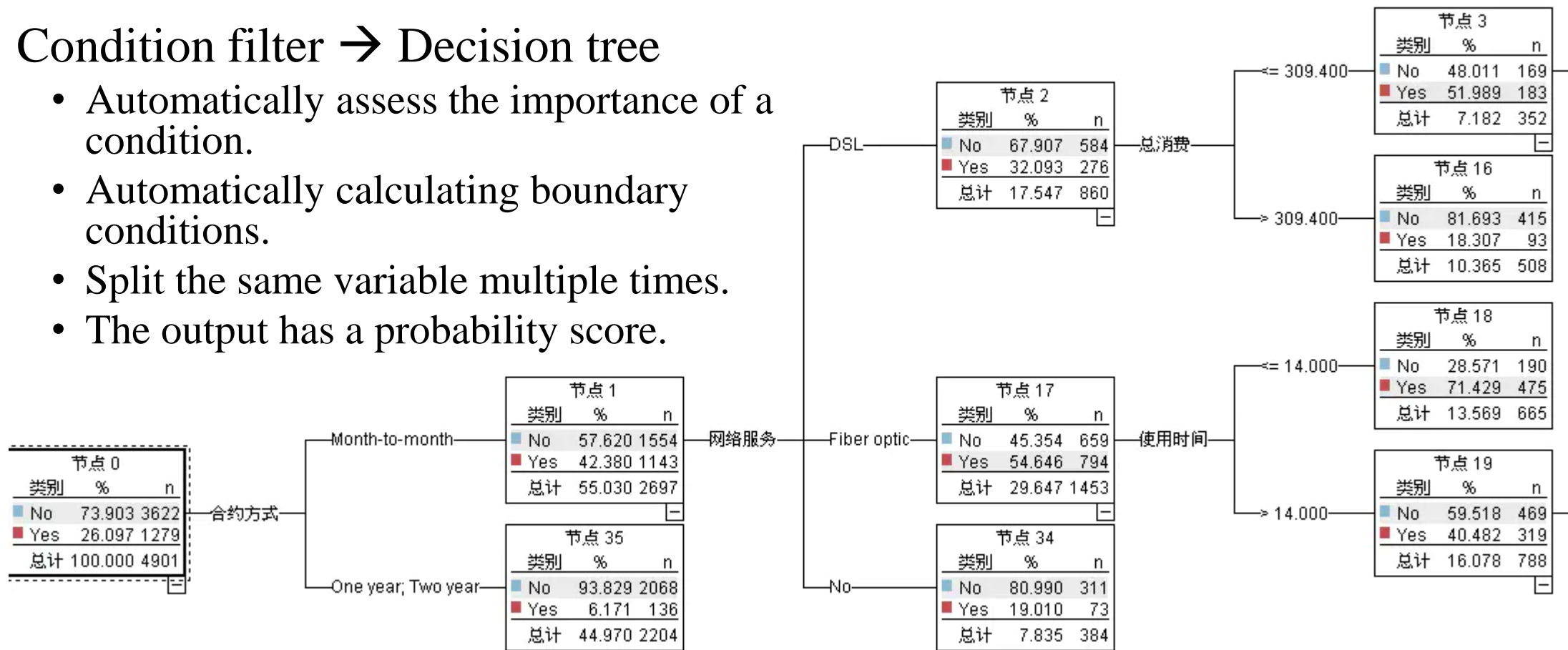
# Shortcomings of the conditional filtering method

- Each of these conditions is a **hard** boundary condition (why 30-50? Is 29-51 OK?). In fact, few conditions have an absolute relationship with the forecast target.
- Multiple criteria cannot be used together (e.g., age 51, but all other criteria meet the requirements extremely well, is it possible to lose?)
- The results are not sorted and the operation is difficult.



# Computers decide the boundaries?

- Condition filter → Decision tree
  - Automatically assess the importance of a condition.
  - Automatically calculating boundary conditions.
  - Split the same variable multiple times.
  - The output has a probability score.



# Scoring alert method

---

$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$   
Customer score = 100 + 5\*Male - 5\*Call\_num - 2\*Call\_time + 10\*Dropped\_num  
If Customer score > 80, the customer churn is warned.

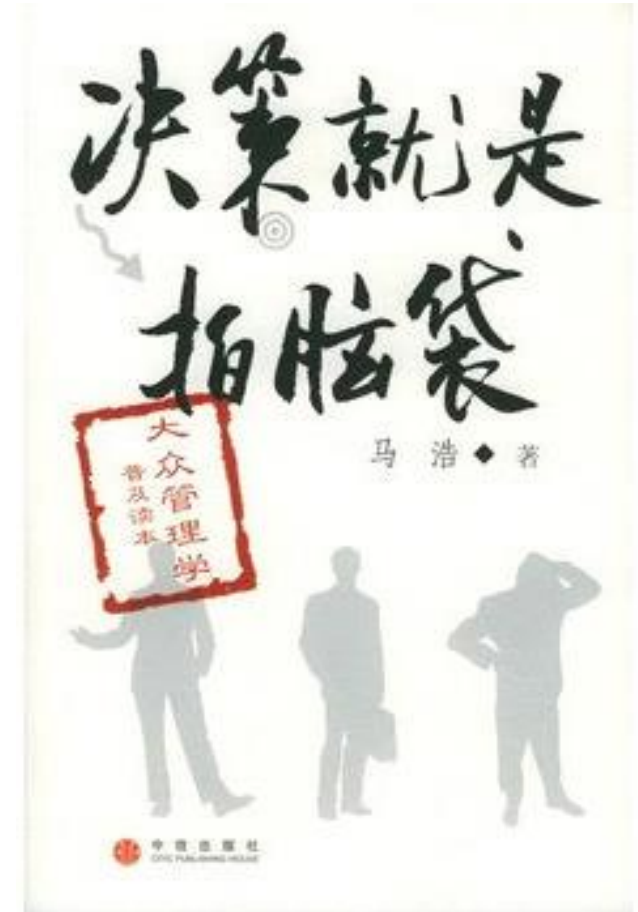
$$Y = \begin{cases} 1 & \text{if } Z > 80 \\ 0 & \text{otherwise} \end{cases} \quad Z = \sum_j \beta_j X_j$$

Issue: Parameters and thresholds are arbitrarily determined;  
controlling the score is challenging.

# Shortcomings of scoring alert method

---

- Weights are entirely arbitrary.
- The weighting of scores lacks scientific justification (why add 20 points for males instead of 19?).
- Scores have no upper limit.
- The non-scientific nature of weights leads to a lack of comparability between scores (for example, a male aged 48, 65 calls, and 320 minutes totaling 60 points; compared to a female aged 45, 1 call, and 15 minutes also totaling 60 points, which is more likely to churn?).
- Because the weights are artificially set, the granularity of scores is often insufficient (for instance, a male aged 48, 65 calls, and 320 minutes totaling 60 points; and a male aged 45, 89 calls, and 480 minutes also totaling 60 points, it's clear that the former is more likely to churn).



# Computers decide the weights?

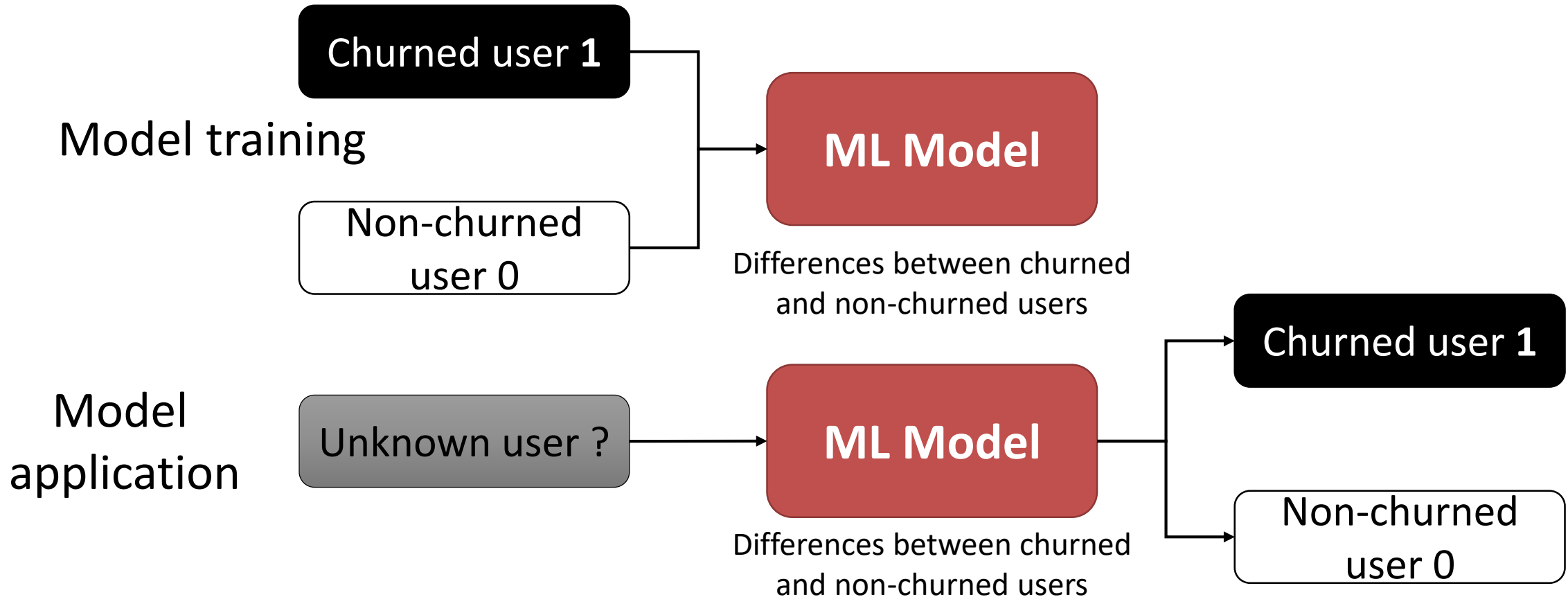
---

- Scoring alert → Logistic Regression
  - Computer decides weights.
  - Scores are between 0 and 1.
  - Scores are probabilities, comparable.

	coef	std err
Intercept	4.7601	0.470
duration	-0.2917	0.015
feton	-1.4144	0.128
gender	1.4394	0.131
call_10086	-0.9040	0.126
peakMinDiff	-0.0024	0.001
edu_class	0.5434	0.078
AGE	-0.0195	0.005
prom	2.3925	0.693
nrProm	-0.7430	0.248
posTrend	-1.5598	0.416
negTrend	-1.3003	0.414
peakMinAv	0.0011	0.000
posPlanChange	-1.0211	0.624

## 4. Training and application of ML

---





# How to select samples?

---

- Target Samples (Positive Samples): The specific objects to be selected for a particular task or purpose.
- Control Samples (Negative Samples): All other ordinary objects apart from the target objects.
- Usually, there should be a relative balance between target and control samples.
- The goal of machine learning is to identify the differences (patterns) between the target and control samples.

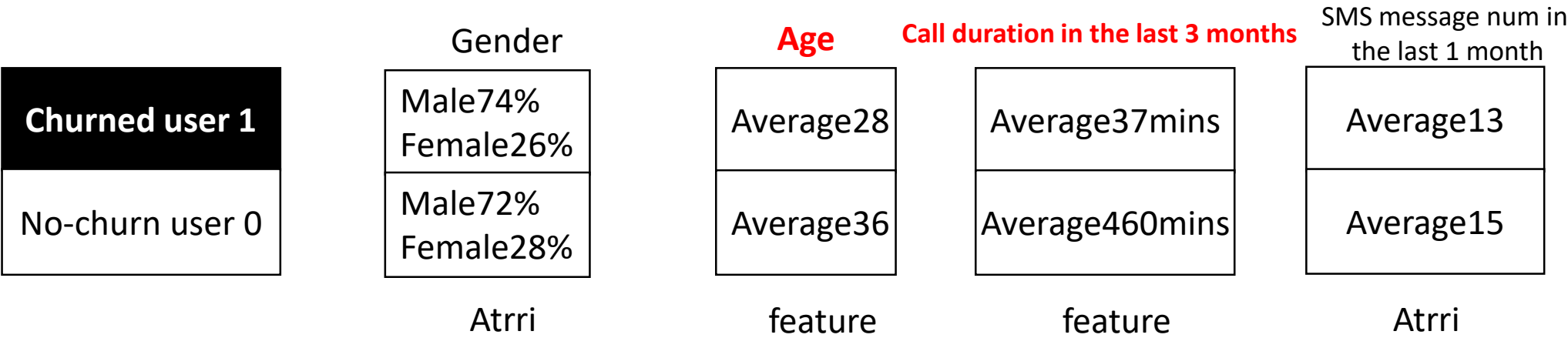
# Describe samples: feature extraction/feature engineering

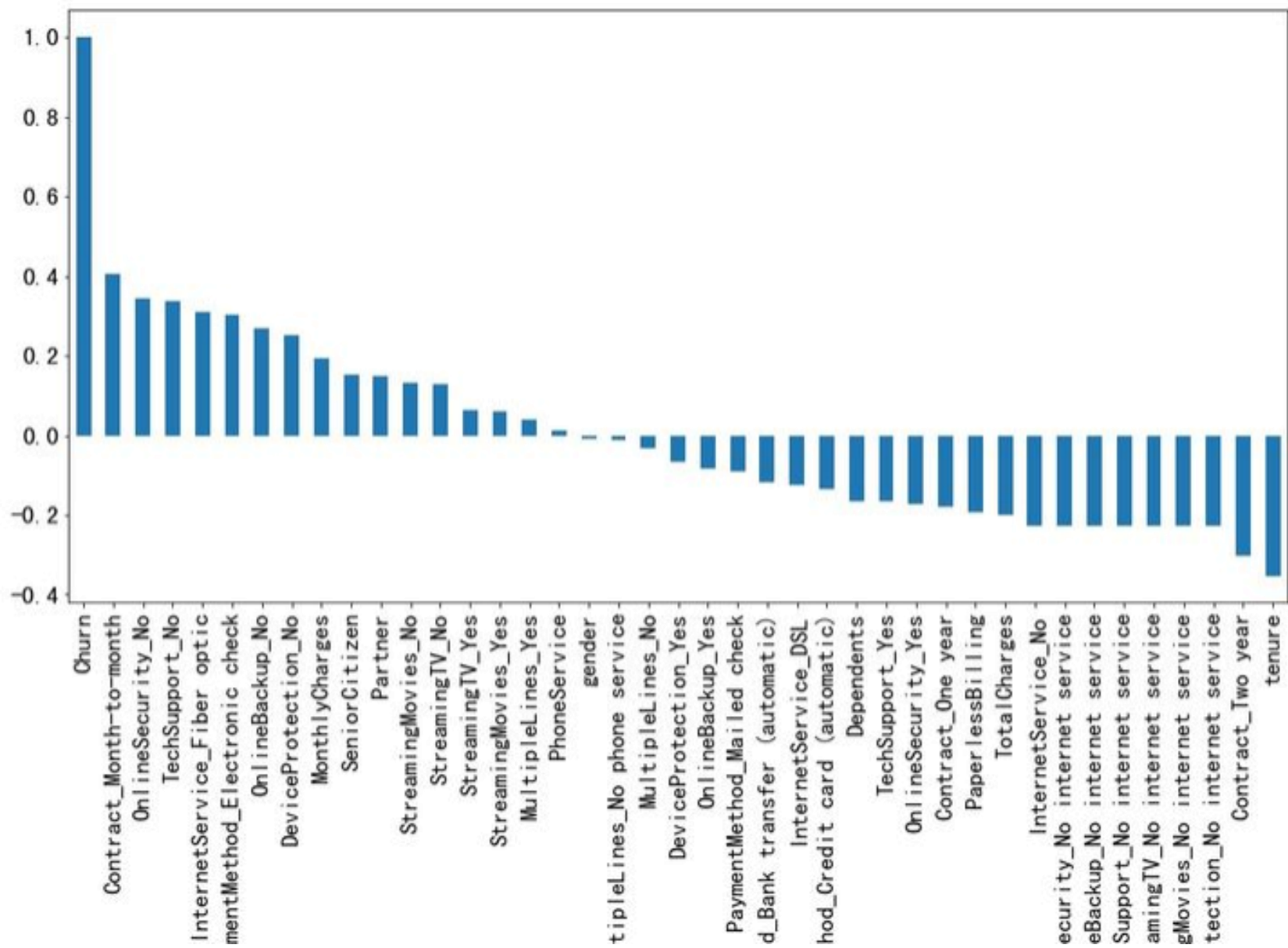
---

- Static Table (1 record per person):
  - User Information Table (profile): User profile information from the previous month, with markers indicating which users have churned.
- Dynamic Tables (number of records per person not fixed) :
  - Call Record Table (cdr\_call): Provides call data for the last six months.
  - Data Record Table (cdr\_mms): Provides data usage details for the last six months.
  - SMS Record Table (cdr\_sms): Provides SMS details for the last six months.
  - Call Drop Table (cdr\_kill): Provides information on call drops in the last six months.
- Based on the dynamic tables, extract attribute data with 1 record per person:
  - the number of calls in the past 6 months, the number of unique communication partners, the count of abnormal call drops, the number of SMS messages in the past 3 months...

Attribute table

Phone number	Sign time	Gender	Age	.....	Call duration in the last 3 months	Call times in the last 3 months	SMS message num in the last 1 month	.....	Churn or not
13...6678	2016-7	Male	28		67	208	34		No
13...2345	2015-3	Male	45		122	466	8		No
13...9854	2020-6	Female	23		89	29	23		Yes





# Deeply Understand "Machine Learning Models"

---

## Model

Choose suitable model  
For example: DT(Decision Tree)、LR(Logistic Regression)、SVM(Support Vector Machine)、NN(Neural Network)、.....

## Feature

Which features?  
Feature number  
Feature types  
For example: Parameters in LR,DT

## Parameter

Training the parameters with samples.  
For example: coefficients in logistic regression, decision order and thresholds in decision tree, etc

Manual setting

Machine computing

- 
- Model selection: Logistic regression  $y = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4+\dots)}}$

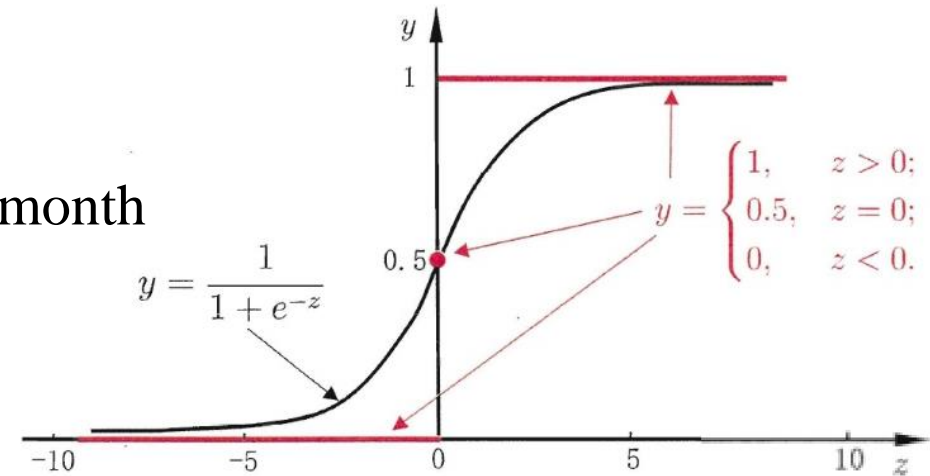
- Feature selection:

- $x_1$ : Age
- $x_2$ : Call duration in the last 3 months
- $x_3$ : Number of SMS messages in the last 1 month
- $x_4$ : Call drop count in the last 3 months
- .....

- Parameter training:

- $\beta_0 = 0.356$ ,  $\beta_1 = 2.403$ ,  $\beta_2 = 3.567$ , .....

- Final Model:  $y = \frac{1}{1+e^{-(0.356+2.403*age+3.567*call+\dots)}}$



$$Y = \frac{1}{1 + e^{-Z}} = \frac{1}{1 + e^{-(8.51 + 0.73 * \text{Male} - 0.25 * \text{Callers} - 0.08 * \text{Call duration} + 6.42 * \text{Call drop count})}}$$

ID	Gender	Callers	Call duration	Call drop count	Z	Y	Prediction Label
185...7203	Male	8	15	2	-8.13	1.000	Churn
186...3204	Female	28	360	4	12.4	0.167	Not churn

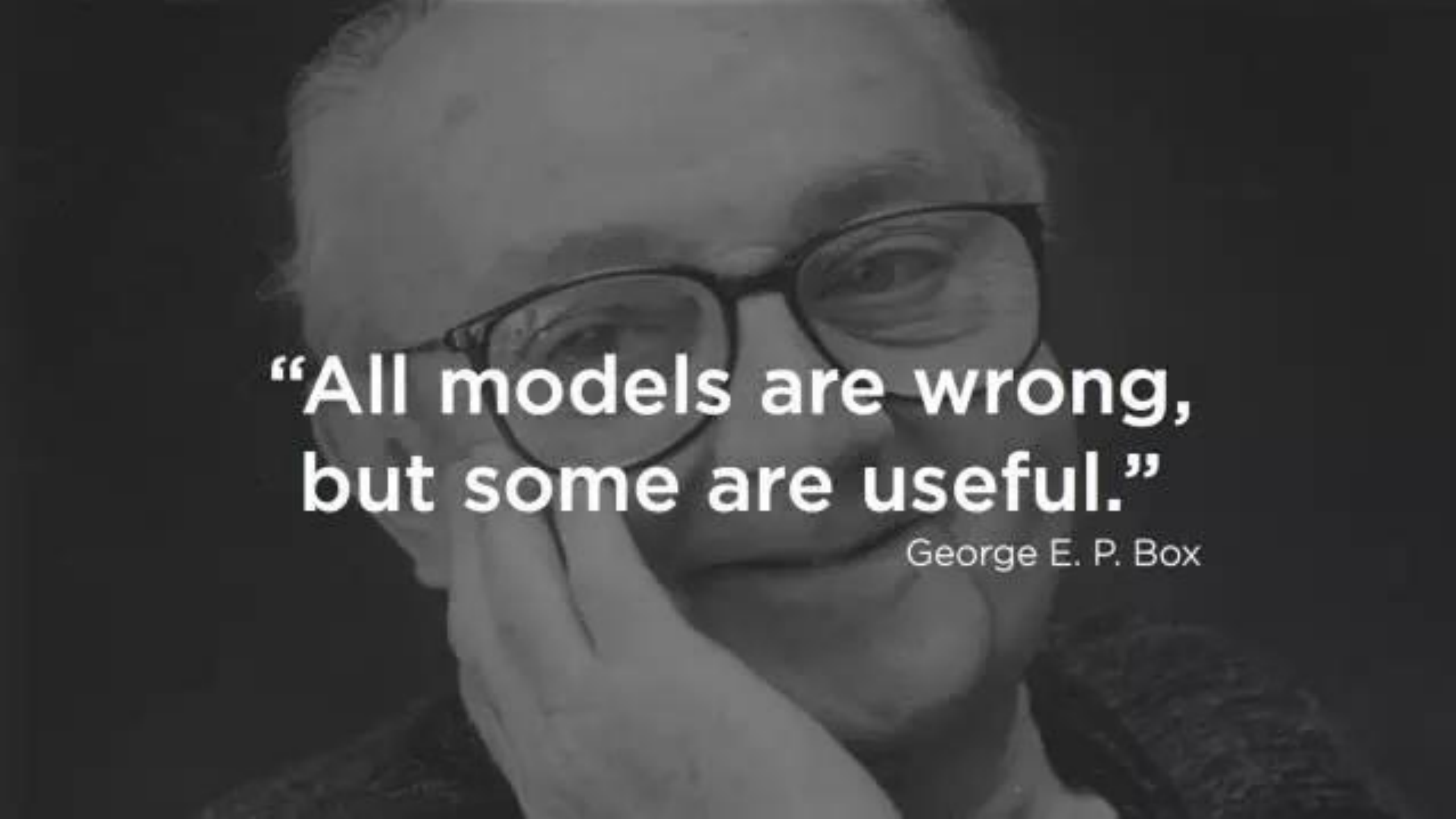
Question: Is this function the "true" function?

The form of this function may not necessarily be the optimal form.

The metrics used may not cover all possible metrics.

Accuracy may not reach 100%.

Humanly understandable, this function is also quite effective.

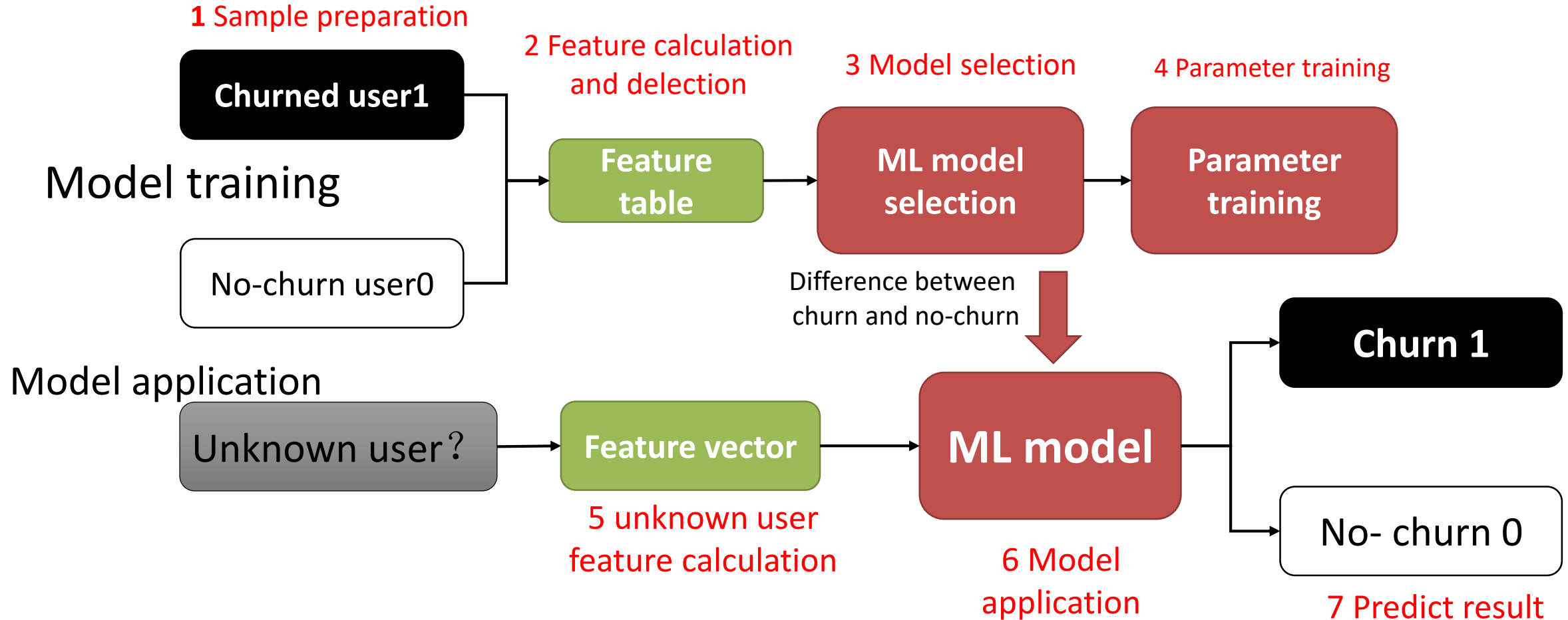


**“All models are wrong,  
but some are useful.”**

George E. P. Box



# Complete process of machine learning



## 2. 数据预处理及特征工程

### Feature table and data preprocessing

# 1.Target of data preprocessing: feature table/wide table

---

- The core object of data preprocessing: the dynamic table or detailed table.
- The processing goal: **another table**, a feature table, or a wide table.
- Isn't the dynamic table already a table?
- Why process it into another table/feature table/wide table?
  - (1) Dynamic tables cannot be directly used for modeling.
  - (2) It's a necessary method for integrating data from multiple sources.

# (1) From one table (dynamic table/detailed table) to another table (feature table/wide table)

---

date	time	Pos	ID	Amount
18-11-23	09:55:13	3	74	3.5
18-11-23	09:56:14	8	76	2.5
18-11-23	09:58:00	11	35	10.0
18-11-23	09:58:11	7	74	5.0
18-11-23	09:59:21	2	32	7.5
18-11-23	09:59:55	3	77	6.0



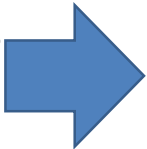
ID	Breakfast times	Average breakfast time	early leaves num	Average breakfast amount	Score
32	18	8:55	8	5.4	75
35	95	7:05	0	5.6	91
74	65	7:40	0	2.7	80
76	34	8:20	2	3.2	69
77	78	6:59	3	4.5	86
79	99	6:55	0	3.5	87

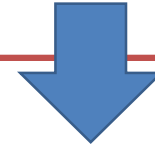
**What's changed?**

**Changes in rows, changes in columns**

# The real basics of data mining

- Another table
- This is often called a feature table or a wide table.

Unit of Analysis  
UoA 



Feature/Attributes

ID	Breakfast times	Average breakfast time	early leaves num	Average breakfast amount	Score
32	18	8:55	8	5.4	75
35	95	7:05	0	5.6	91
74	65	7:40	0	2.7	80
76	34	8:20	2	3.2	69
77	78	6:59	3	4.5	86
79	99	6:55	0	3.5	87

ID

Input

Target

# (2) 多来源数据的融合

## Fusing data from multiple sources

ID	早餐次数 breakfast times	平均早餐时间 Ave breakfast time	早退次数 early leaves times	平均早餐金额 Ave break amount	超市购物金额 shop amount	图书馆 借阅数 books borrow	洗澡时间熵 Entropy of bath time	性别 Gender	Score
32	18	8:55	8	5.4	453	6	1.5	男M	75
35	95	7:05	0	5.6	24	34	2.0	女F	91
74	65	7:40	0	2.7	88	7	1.1	男M	80
76	34	8:20	2	3.2	789	23	3.5	女F	69
77	78	6:59	3	4.5	43	77	0.8	男M	86
79	99	6:55	0	3.5	86	0	3.1	男F	87

就餐数据  
500→1

购物  
50→1

借阅  
n→1

洗澡  
10→1

基本  
信息1

成绩  
1

# 直接的数据关联会导致局部笛卡尔积

## Direct data association leads to a local Cartesian product

---

ID	性别 Gender	成绩 Score
32	男M	75

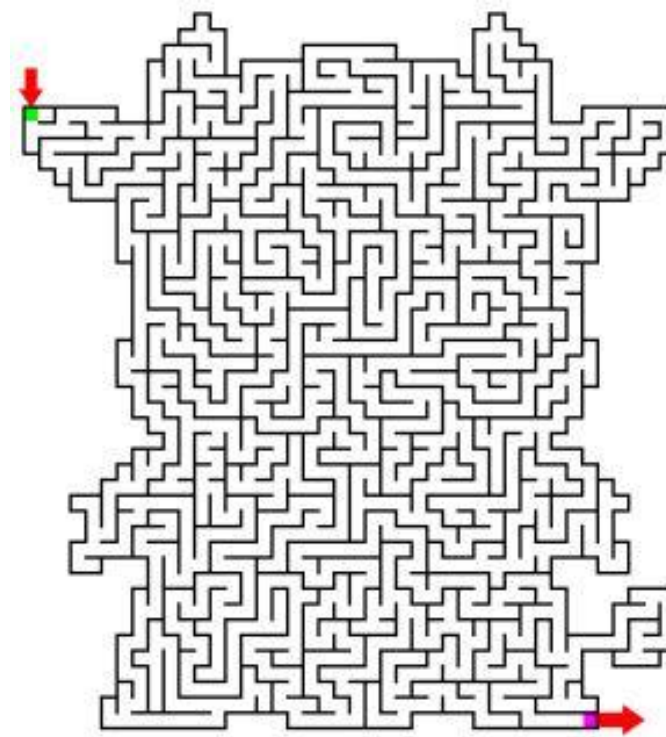
ID	性别 Gender	金额 Amount
32	男M	75
32	女F	91
32	男M	80
32	女F	69
32	男M	86
32	男F	87

ID	性别 Gender	成绩 Score
32	男M	75
35	女F	91
74	男M	80
76	女F	69
77	男M	86
79	男M	87

## 2. 构建特征表：业务主导、谋事在前

### Building the feature table: Business-driven, planning in front.

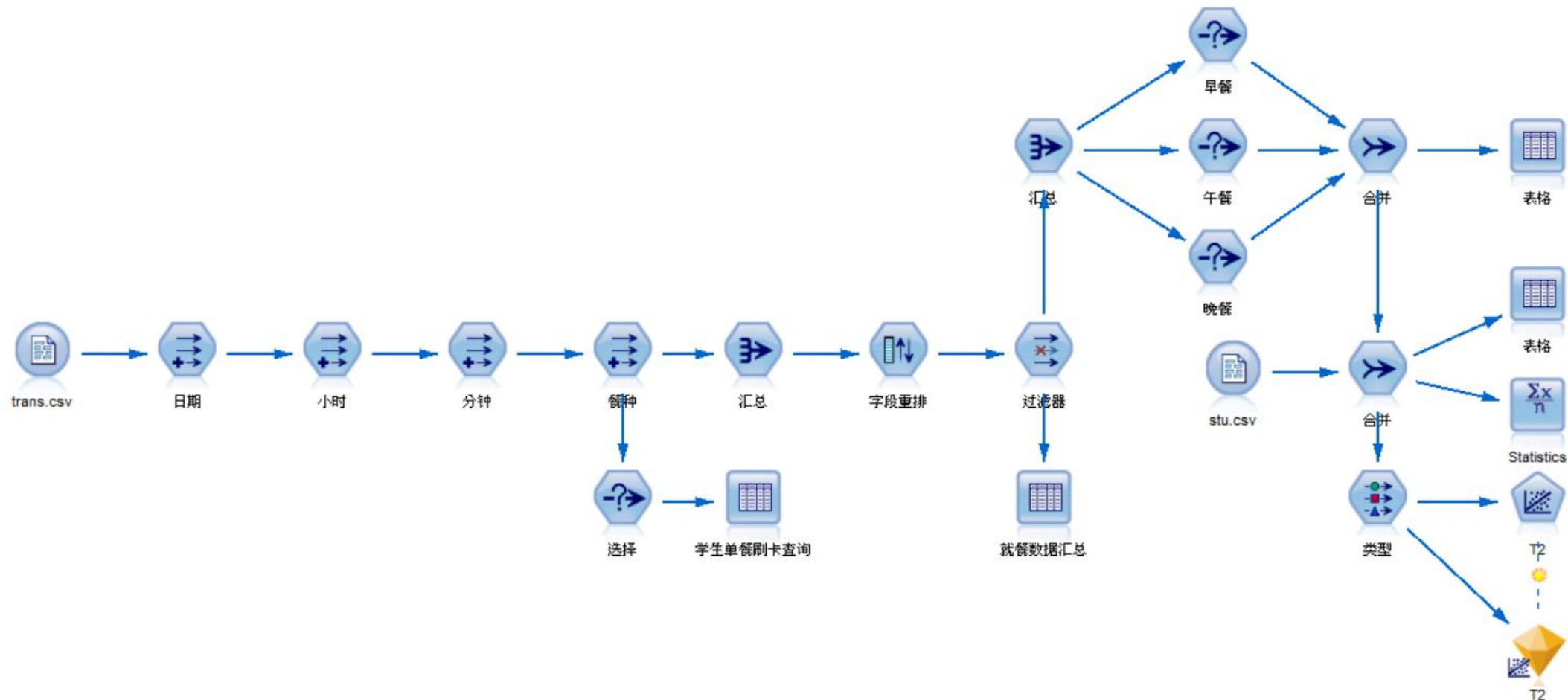
- Step 1: Define the prediction target
- Step 2: Clearly analyze the object UoA
- Step 3: Think about what features are needed  
(business driven)
- Step 4: thinking process  
(to get simple example by hand)
- Step 5: Get started





### 3. 案例：学生就餐行为特征表

### 3. Case: Feature table for student dining behavior.



1.详细的就餐数据：1人1天1餐多次刷卡多条数据，共计：9.3万条  
Detailed dining data: Multiple entries per day per person, totaling 93,000 entries.

stuid	campus	canteen	pos	transtime	transvalue
32	北	2	112	2014-10-31 07:34:31	2.000
32	北	2	100	2014-10-31 11:49:55	1.000
35	北	2	112	2014-10-31 16:21:01	2.100
46	北	2	65	2014-10-31 16:48:18	8.000
52	北	2	65	2014-10-31 16:48:15	8.000
54	北	2	65	2014-10-31 16:48:53	8.000
55	北	2	100	2014-10-31 18:36:40	1.000
56	北	2	112	2014-10-31 16:20:47	2.800
61	北	2	112	2014-10-31 07:44:49	2.000
72	北	2	89	2014-10-30 11:40:52	6.000
73	北	2	112	2014-10-30 07:38:38	1.500



1餐多次刷卡合并为1条  
Combining multiple card swipes for one meal into a single entry.

2.压缩就餐数据：1人1天1餐1条数据，共计：4.9万条  
Compressing dining data: One entry per person per day per meal, totaling 49,000 entries

stuid	日期	餐种	刷卡次数	transvalue_Sum	分钟_Min
32	2014-10-31	早餐	2	3.600	454
32	2014-10-31	午餐	2	8.000	708
35	2014-10-31	晚餐	3	8.100	971
46	2014-10-31	晚餐	2	10.000	1008
52	2014-10-31	晚餐	3	11.400	1004
54	2014-10-31	晚餐	4	17.400	1005
55	2014-10-31	晚餐	2	5.000	1116
56	2014-10-31	晚餐	2	12.800	969
61	2014-10-31	早餐	2	5.000	464
72	2014-10-30	午餐	2	7.500	700
73	2014-10-30	早餐	4	7.500	449

1人1种餐合并为1条  
Combining one person's one type of meal into a single entry.



4.合并就餐数据：1人1条数据，共计：107条  
Combining dining data: One entry per person, totaling 107 entries.

ID	早餐次数	平均早餐金额	午餐次数	平均午餐金额	晚餐次数	平均晚餐金额
1	146	3.241	200	10.247	176	9.221
2	2	3.400	42	13.586	34	10.191
3	158	2.131	243	5.223	233	5.840
4	238	2.597	272	6.455	252	6.284
5	126	2.448	192	7.487	198	8.192
6	135	2.216	228	6.120	223	5.853
7	111	3.622	184	7.829	186	8.657
8	56	4.470	132	10.364	115	10.607
9	109	1.932	149	7.476	140	7.393
10	193	1.399	232	4.440	233	3.842

1人3种餐合并为1条



Combining one person's three types of meals into a single entry

3.压缩就餐数据：1人1餐1条数据，共计：107\*3条  
Compressing dining data: One entry per person per meal, totaling 107 x 3 entries.

ID	餐种	就餐次数	平均就餐金额	平均就餐时间	平均刷卡次数
1	早餐	146	3.241	451.164	1.055
1	午餐	200	10.247	702.995	1.985
1	晚餐	176	9.221	1064.358	2.034
2	早餐	2	3.400	435.500	1.500
2	午餐	42	13.586	711.452	2.714
2	晚餐	34	10.191	1078.029	2.412
3	早餐	158	2.131	469.804	1.234
3	午餐	243	5.223	707.872	1.362
3	晚餐	233	5.840	1076.670	2.197
4	早餐	238	2.597	470.017	1.168
4	午餐	272	6.455	718.298	1.294
4	晚餐	252	6.284	1069.536	1.389

合并就餐数据：1人1条数据，共计：107条  
Combining dining data: One entry per person, totaling 107 entries.

ID	早餐次数	平均早餐金额	午餐次数	平均午餐金额	晚餐次数	平均晚餐金额
1	146	3.241	200	10.247	176	9.221
2	2	3.400	42	13.586	34	10.191
3	158	2.131	243	5.223	233	5.840
4	238	2.597	272	6.455	252	6.284
5	126	2.448	192	7.487	198	8.192
6	135	2.216	228	6.120	223	5.853
7	111	3.622	184	7.829	186	8.657
8	56	4.470	132	10.364	115	10.607
9	109	1.932	149	7.476	140	7.393
10	193	1.399	232	4.440	233	3.842



个人基础信息：1人1条数据，共计：107条Personal basic information: One entry per person, totaling 107 entries.

ID	Gender	T1	T2	Class	Group	Prov	City	BirthDay
1	1	88	83	1	1	福建	龙岩市	1995-02-01
2	1	76	69	1	1	福建	龙岩市	1996-05-01
3	1	75	74	1	1	福建	龙岩市	1996-10-01
4	1	90	91	2	1	北京	朝阳区	1995-11-12
5	1	86	86	2	1	江西	景德镇市	1996-01-13
6	1	88	91	3	1	内...	呼和浩特市	1996-02-21
7	1	82	75	3	1	广东	湛江市	1995-09-21
8	1	76	83	3	1	辽宁	大连市	1997-10-21
9	1	79	83	3	1	辽宁	大连市	1996-02-22
10	1	74	90	3	1	浙江	杭州市	1995-09-22

最终特征表：1人1条数据，共计：107条Final feature table: One entry per person, totaling 107 entries.

ID	BirthDay	Gender	Prov	City	Class	Group	早餐次数	平均早餐金额	午餐次数	平均午餐金额	晚餐次数	平均晚餐金额	T1	T2
1	1995-02-01	1	福建	龙岩市	1	1	146	3.241	200	10.247	176	9.221	88	83
2	1996-05-01	1	福建	龙岩市	1	1	2	3.400	42	13.586	34	10.191	76	69
3	1996-10-01	1	福建	龙岩市	1	1	158	2.131	243	5.223	233	5.840	75	74
4	1995-11-12	1	北京	朝阳区	2	1	238	2.597	272	6.455	252	6.284	90	91
5	1996-01-13	1	江西	景德...	2	1	126	2.448	192	7.487	198	8.192	86	86
6	1996-02-21	1	内...	呼和...	3	1	135	2.216	228	6.120	223	5.853	88	91
7	1995-09-21	1	广东	湛江市	3	1	111	3.622	184	7.829	186	8.657	82	75
8	1997-10-21	1	辽宁	大连市	3	1	56	4.470	132	10.364	115	10.607	76	83
9	1996-02-22	1	辽宁	大连市	3	1	109	1.932	149	7.476	140	7.393	79	83
10	1995-09-22	1	浙江	杭州市	3	1	193	1.399	232	4.440	233	3.842	74	90

### 3. 机器学习模型的种类及应用场景

### Types of machine learning models and their application scenarios



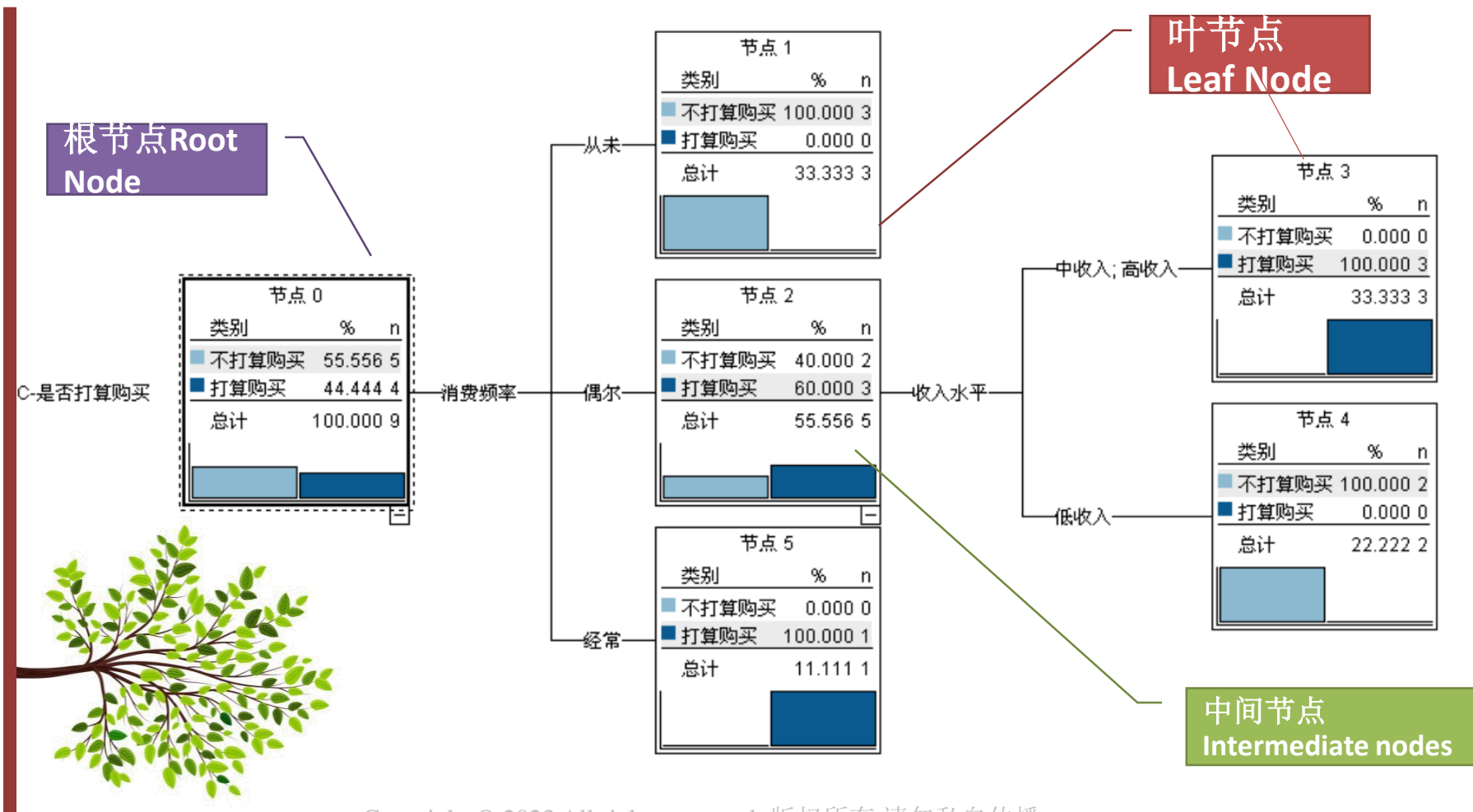
# 1. 传统机器学习模型

## Traditional machine learning models

---

- 有监督学习Supervised learning
  - 分类Classification（预测标签predict label）
    - **逻辑回归Logistic Regression**
    - **决策树 Decision Tree**、随机森林 Random Forest
    - 支持向量机SVM
  - 回归Regression（预测数值predict value）
    - **线性回归Linear Regression**
    - 支持向量回归SVC、分类回归树Classification regression tree
- 无监督学习Unsupervised learning
  - Relationships between objects
    - clustering
    - Outlier detection
  - Relationships between variables
    - Association rules
    - Dimensionality reduction

## 2. Decision Tree





- 
- The results of its analysis are presented in a way that looks like an inverted tree.
  - It reflects the continuous grouping process of sample data.
  - Decision trees are divided into Binary trees, classification trees, and regression trees.
  - It reflects the logical relationship between the values of input variables and output variables.
  - Logical comparison formally expresses a kind of inference rule.
  - Each leaf node corresponds to an inference rule.
  - Classification prediction for new data objects.

- 常用分类器  
Common classifiers
- Wide application
  - Fraud detection
  - Disease diagnosis
  - Marketing
  - Accident analysis

相似文献

## 决策树模型与logistic回归模型在胃癌高危人群干预效果影响因素分析中的应用

目的 采用决策树模型与logistic回归模型分析影响农村胃癌高危人群干预效果的影响因素.方法 根据胃癌高危人群干预效果及其相关因素,分别建立决策树模型和logistic回归...

刘兵, 李苹, 朱玫烨, ... - 《中国卫生统计》

被引量: 7 发表: 2018年

## Logistic回归、决策树和神经网络在预测2型糖尿病并发末梢神经病变中的性能比较

近年来,数学方法和计算机技术的发展使复杂的模型预测成为可能.目前能够建立预测模型的方法主要有统计学方法和数据挖掘方法,基于这两类方法的预测技术已逐渐被应用在生...

李长平 - 《中国人民解放军军事医学科学院》

被引量: 7 发表: 2009年

## 决策树模型与logistic回归在中学生尝试吸烟影响因素中的应用

目的 了解江苏省中学生吸烟的分布情况及影响因素,为针对性地开展中学生控烟干预提供科学依据.方法 本研究数据来源于2013年江苏省青少年烟草调查,采用多阶段分层整群随...

曲晨, 覃玉, 毛涛, ... - 《中国慢性病预防与控制》 2020年28卷4期 264-269页 ISTIC PKU CA

被引量: 0 发表: 2020年

## Logistic回归和决策树在数据库营销响应中的应用

与传统营销方式相比,数据库营销不但能提高营销效益,而且是客户关系管理的基础,是企业从以产品为中心向以客户为中心的经营体系转型的杠杆.由于强大的数据存储和挖掘功...

冯伟 - 《兰州财经大学》



# Overview of the decision tree algorithm

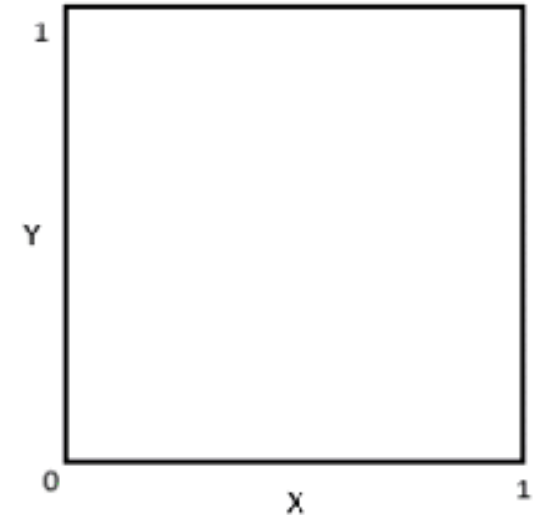
---

- The process of building a decision tree is the process of forming each branch of the decision tree in turn.
- Each branch of the decision tree completes the region division of the  $n$ -dimensional feature space under certain rules.
- Once the tree is built, the  $n$ -dimensional feature space is divided into small rectangular regions with boundaries parallel to or perpendicular to the axes.

# Decision tree generation logic

---

- Generation/Growth
- When determining the criteria for dividing the feature space at each step, take into account both regions that will be formed. The goal is to ensure that the sample points contained in the two resulting regions are as "pure" as possible simultaneously.



For more tutorials: [annalysin.wordpress.com](http://annalysin.wordpress.com)

# Decision tree steps

---

**Grow**



**Prune**

Use of the **training sample** set  
to complete the  
establishment of a decision  
tree process

Using the **test sample** set to  
streamline formed by decision  
tree

A model that performs better during training is likely to  
perform worse during application

# Model Scaling: Ensemble Learning

---

- Two heads are better than one
- The presence of bias and variance often leads to a lack of robustness in predictions made by a model built on a training dataset.
- Strategies in data mining:
  - Bagging technique
  - Boosting technique
- Both involve modeling and voting stages.

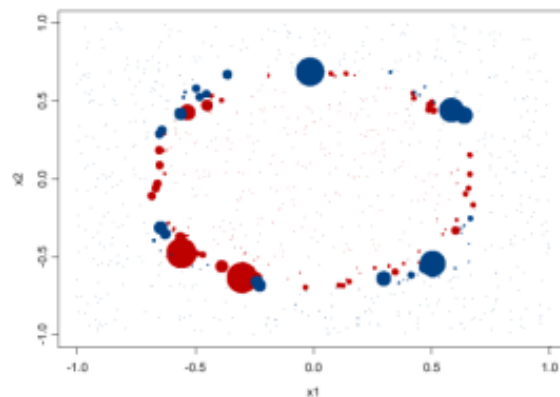
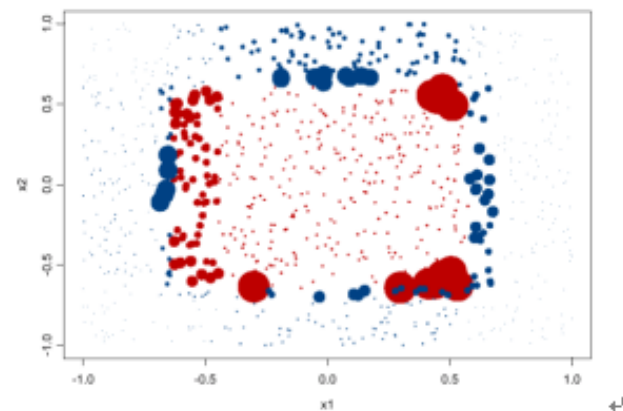
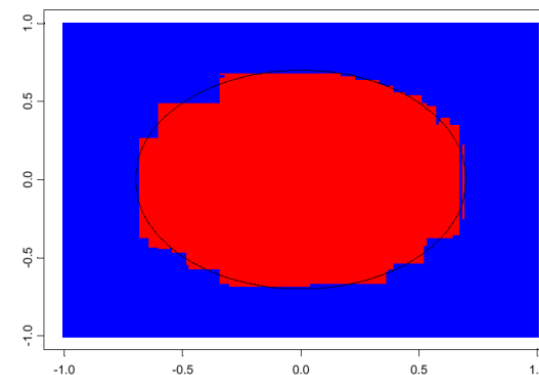
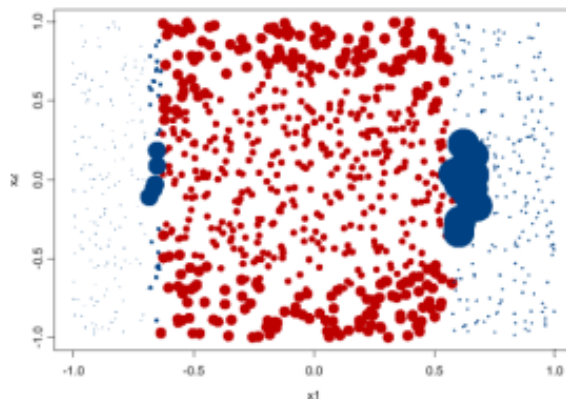
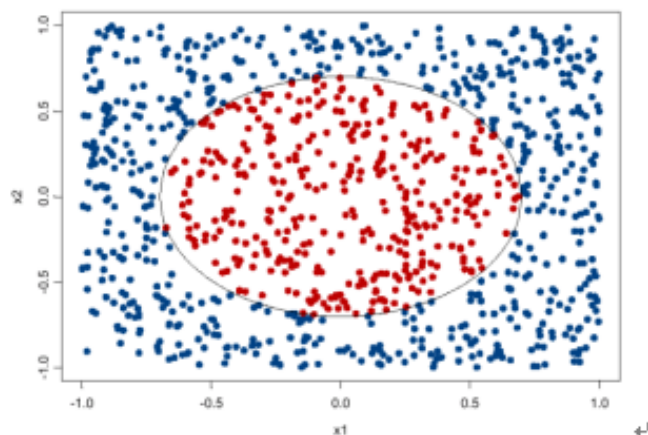
# Bagging

---

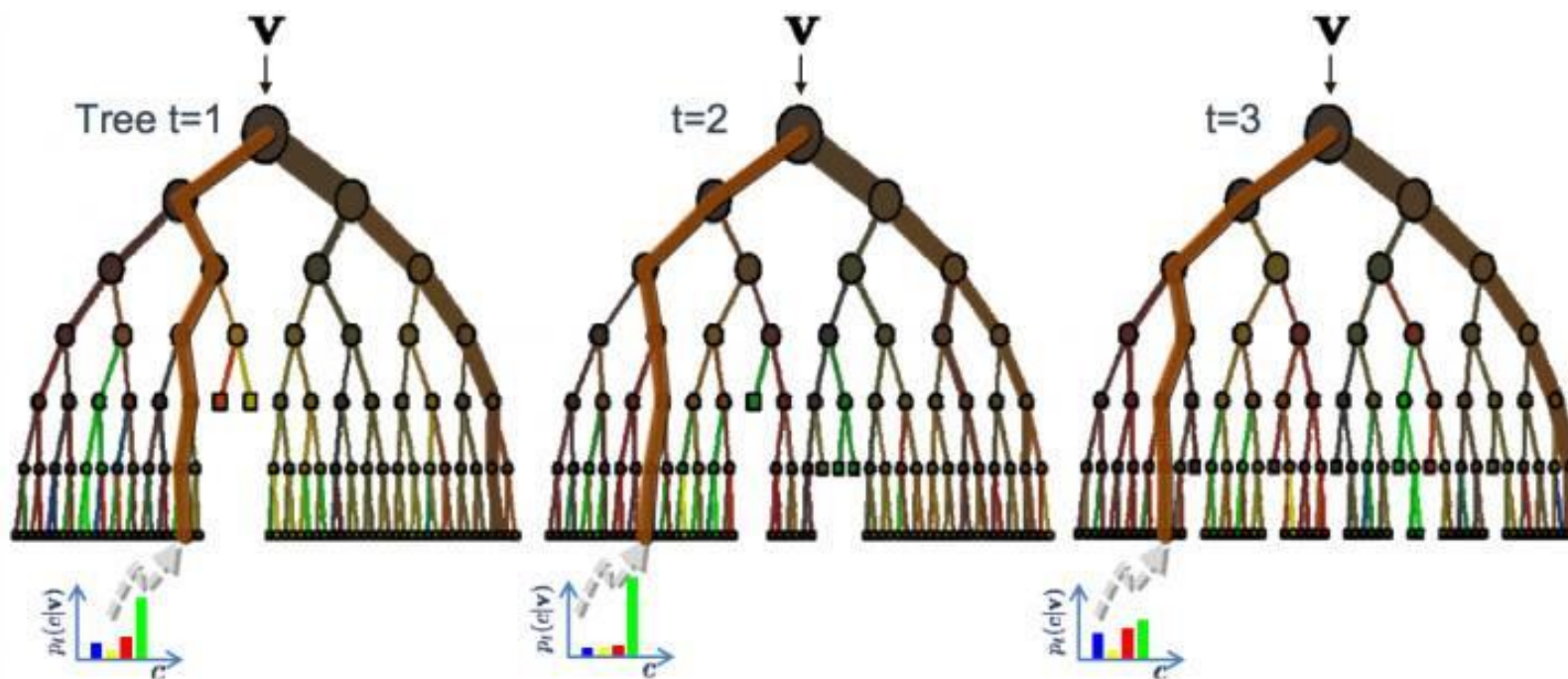
- Modeling process(Input: training sample set  $T$ , training number  $k$ ;  
Output: Multiple decision tree models  $C_1, C_2, \dots$ )
- For  $i=1, 2, \dots, k$  do
  - Samples are randomly drawn from  $T$  with replacement to form a sample set  $T_i$  with the same sample size
  - The model  $C_i$  is constructed using  $T_i$  as the training set
- End for

# Boosting技术

- 建立k个模型Build k models.; k个模型投票The k models vote

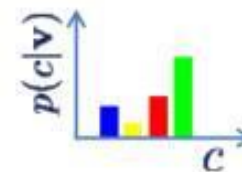


# 随机森林 (Random Forest)



The ensemble model

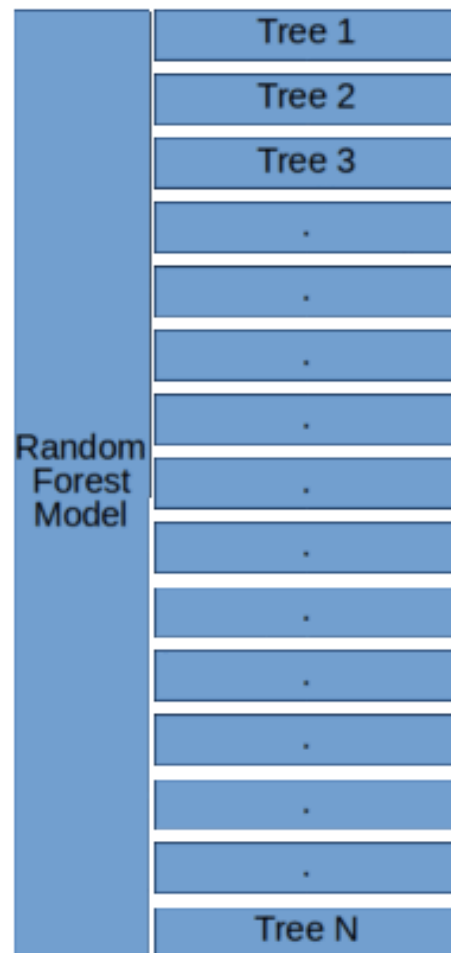
$$\text{Forest output probability } p(c|\mathbf{V}) = \frac{1}{T} \sum_t^T p_t(c|\mathbf{V})$$



# 随机森林的决策结果

## Decision results from random forests

---

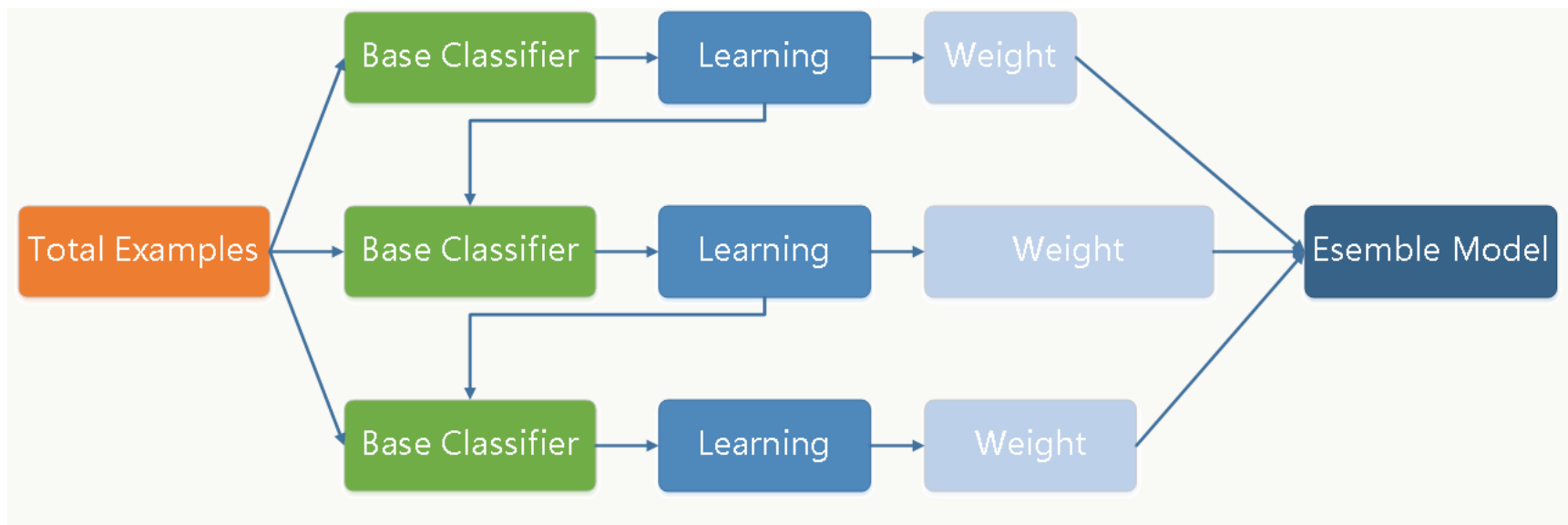




# GBDT和XGBoost

---

- 随机森林是目前商业领域运用最为广泛的一种算法。Random forest is one of the most widely used algorithms in business.
- 特别是模型算法竞赛领域。Especially in the area of model algorithm competitions.
- GBDT (Gradient Boosting Decision Tree), XGBoost (eXtreme GBoost)



### 3. Regression Analysis

---

- Regression analysis:
  - A weighted sum of some variables is used to predict the target variable
  - The scientific method of integration
  - Example: the possibility of churn
  - Integral method: opened a service, the product x points; Call volume, product x points
  - Regression method: Likelihood of churn =  $f(\text{coefficient } 1 * \text{whether X service has been opened} + \text{coefficient } 2 * \text{call volume})$
- General linear regression analysis: The target variable is continuous
- Logistic Regression (LR) : target variable 0-1

# 客户价值预测（数量预测）

## Customer value Prediction (quantity prediction)

---

每个客户有不同的客户价值  
Each customer has a different customer value

模型训练  
Model training

老客户数据  
Old customer data

特征表  
Feature Table

机器学习  
模型选择  
Model selection

参数训练  
Parameter train

各种特征与  
客户价值的  
关系

The relationship between  
various characteristics and  
customer value

模型应用  
Model Application

新客户  
New customer

特征向量  
Feature Vector


机器学习  
模型  
ML Model


客户价值  
Customer value


# 一般线性回归General linear regression


---


$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$


 客户购买量  
Customer purchase quantity

 基准  
Baseline

 性别  
Gender

 月收入  
Monthly income

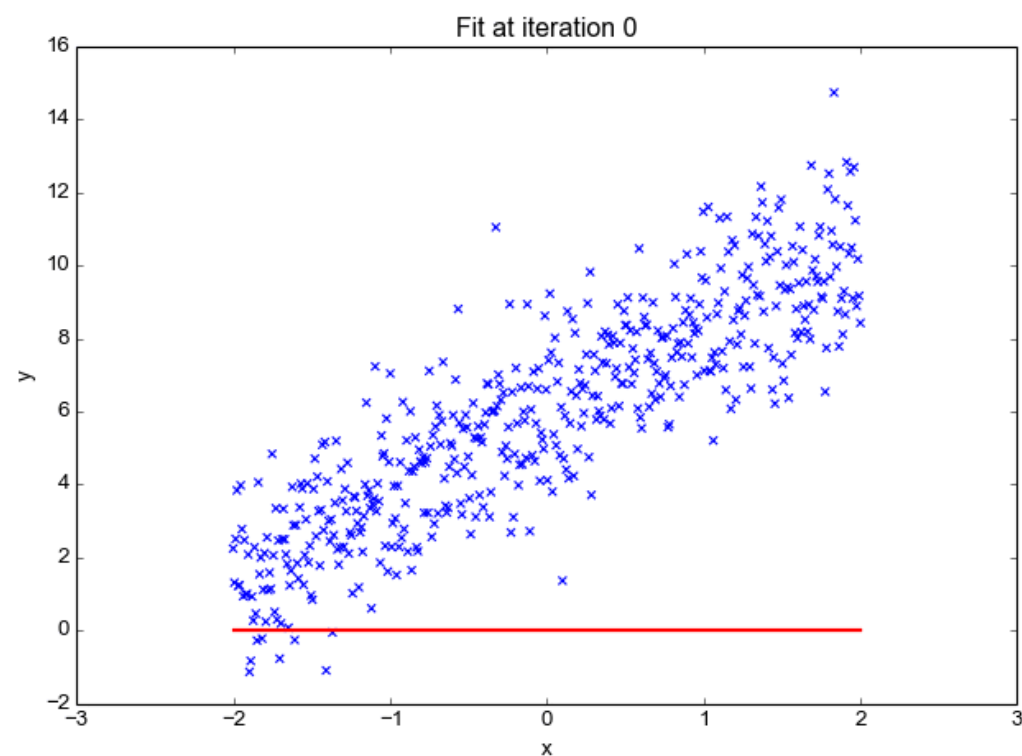
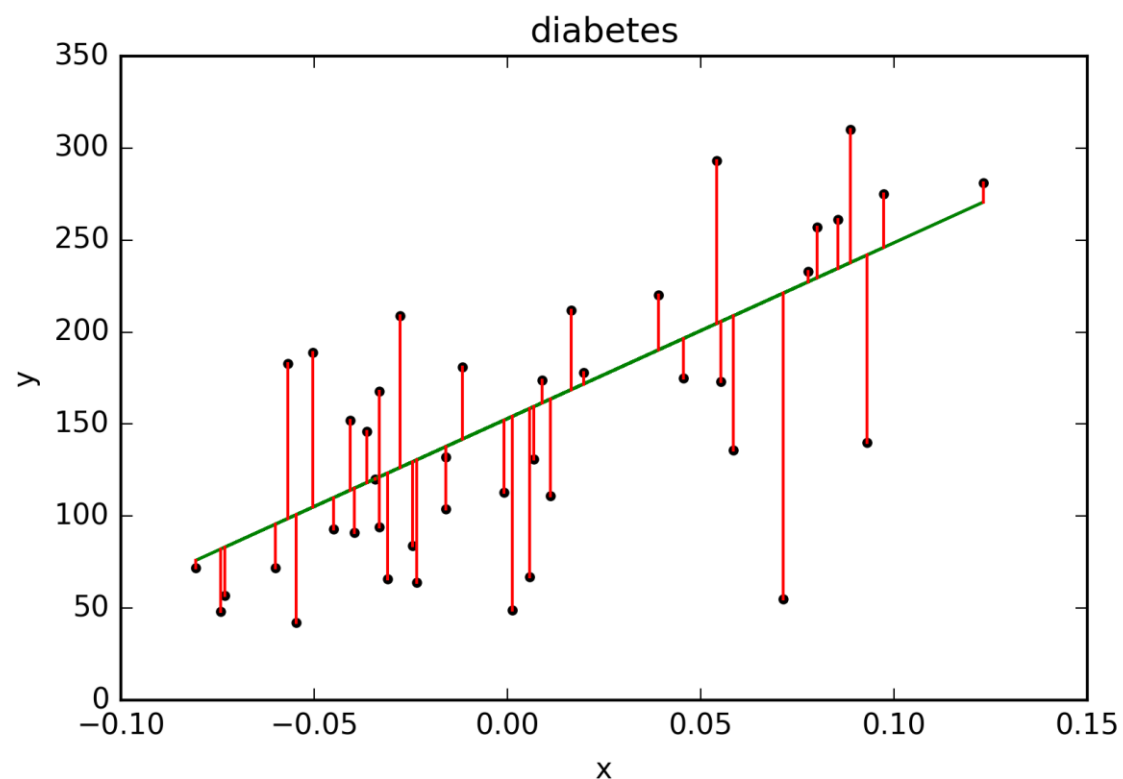
 历史购买量  
Historical purchase volume

 其他  
else

# 最小二乘法Least squares method

---

- 显式解 及 梯度下降法Explicit solutions and gradient descent methods

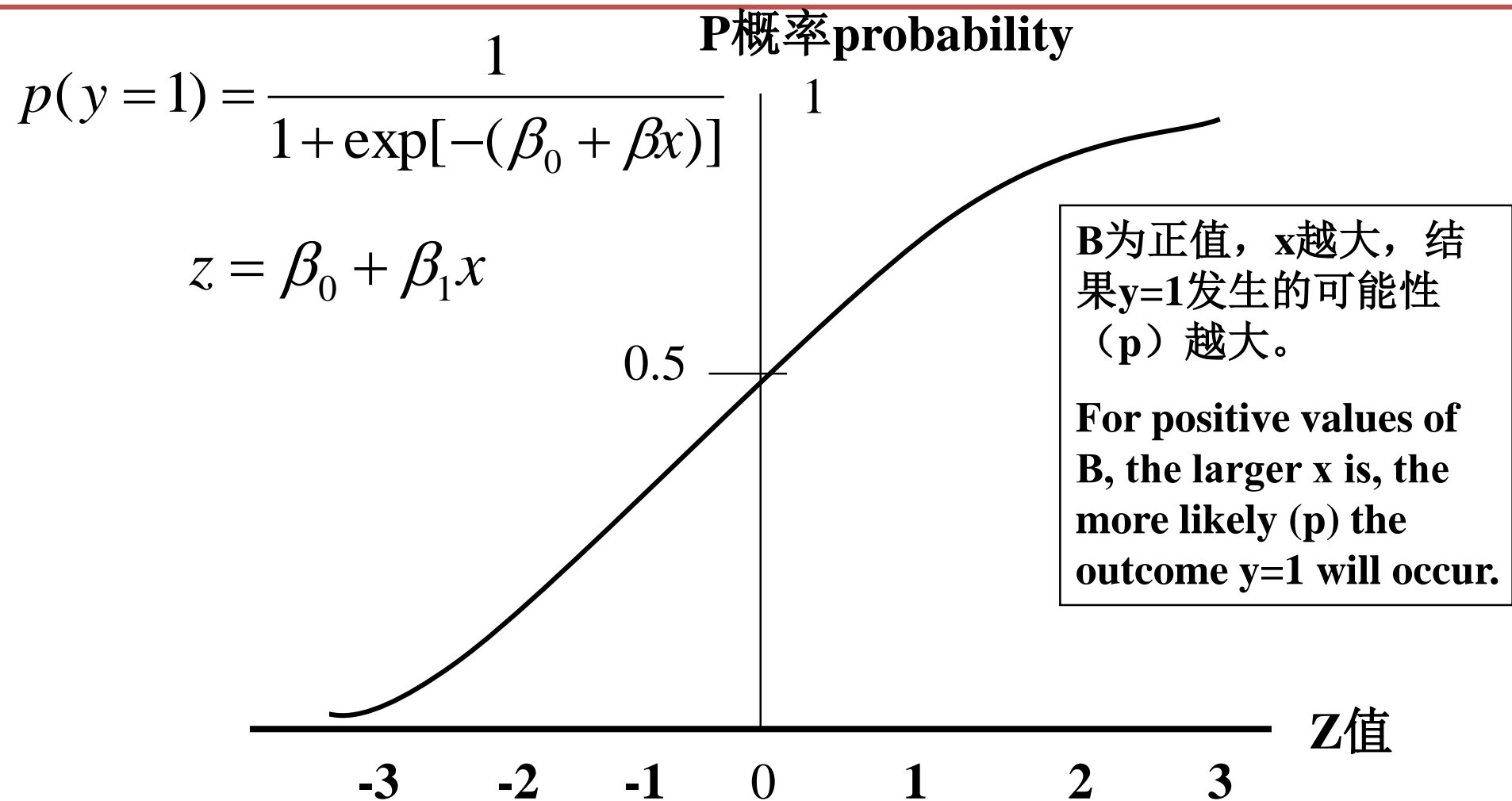


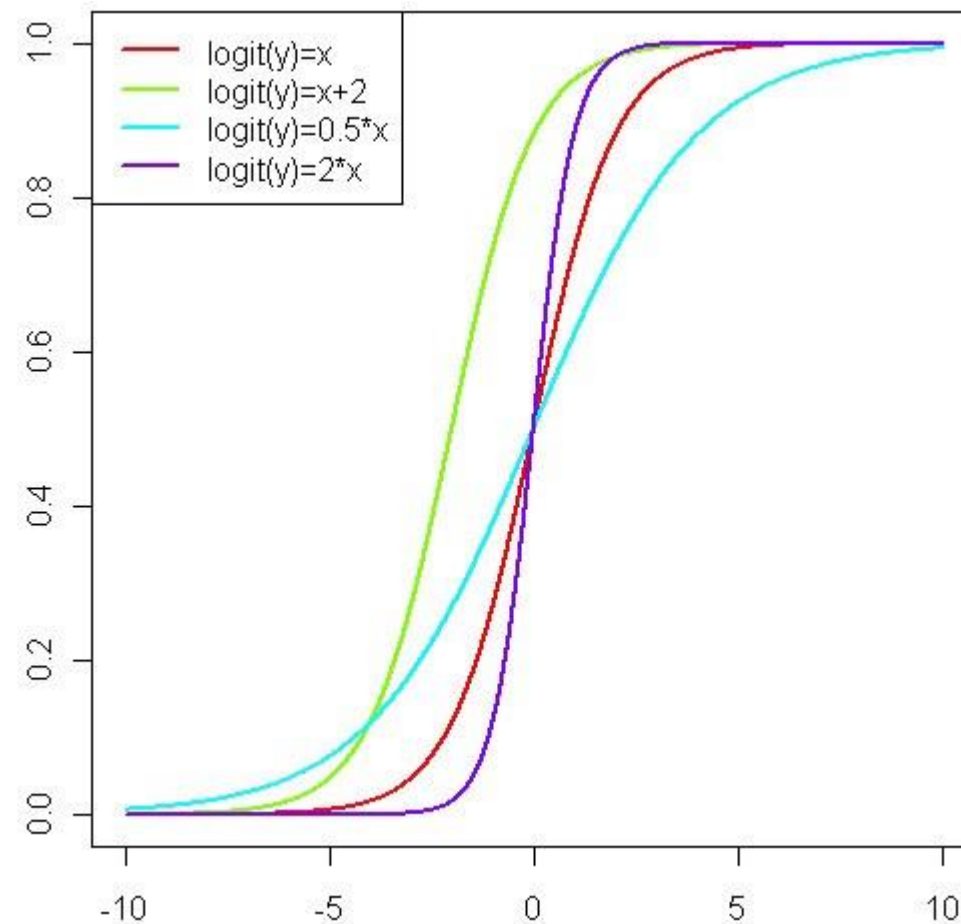
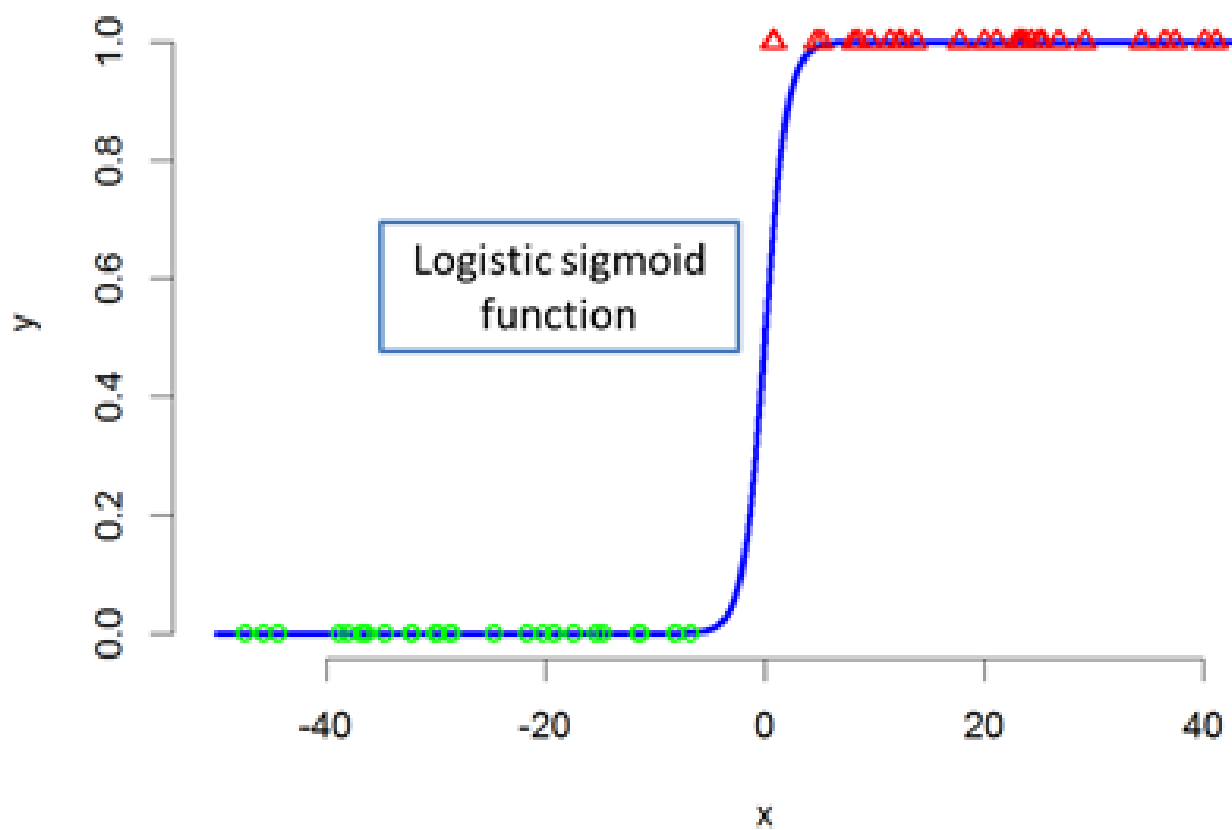
# 回归分析结果举例：P值与R方

## Examples of regression results: P-values and R-squares

变量名feature	估计值Estimated value	P-value	变量名feature	估计值Estimated value	P-value
常数项constant	-0.162	<0.001	当期保养总花费 Current maintenance cost	0.258	<0.001
车型Car type-A	0.001	0.067	当期保养总次数 Current number of maintenance	0.129	<0.001
车型Car type-B	0.260	<0.001	当期新增里程数 Current New mileage	0.198	<0.001
车型Car type-C	0.113	0.159	累积购车数量 Accumulate the number of vehicles purchased	0.055	<0.001
车型Car type-D	0.176	0.035	车价 Car price	0.143	<0.001
车型Car type-其它else	-0.094	0.273			
调整R方(Adjusted R-squared): 36.37%					

# Logistic回归 逻辑回归 逻辑斯迪/蒂克回归







# Calculation of the predicted value

---

- Calculation formula:
- The score is a score between 0 and 1 
$$p(y = 1) = \frac{1}{1 + \exp[-(\beta_0 + \beta x)]}$$
- Score =  $1/(1 + \exp(-\text{final score}))$
- A score  $>0.5$  is judged to be 1, and the probability score is the original score
- If the score is  $<0.5$ , the decision is 0, and the probability score is  $1 - \text{original score}$

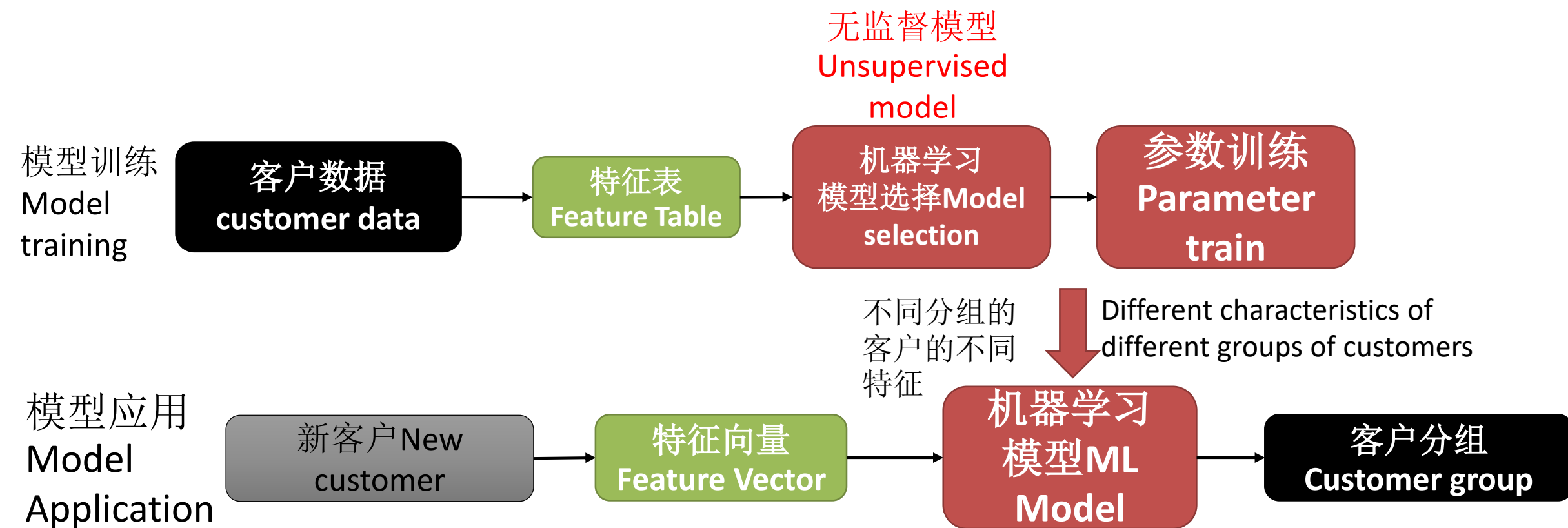
## 4. Cluster analysis

---

- Cluster analysis is a method of describing data to build a model, and the purpose is to explore whether there is a "natural subclass" in the data.
- Cluster analysis: The records are clustered such that records of the same category are as similar as possible to each other and records of different categories are as different as possible.

# 客户分组研究（没有标签）

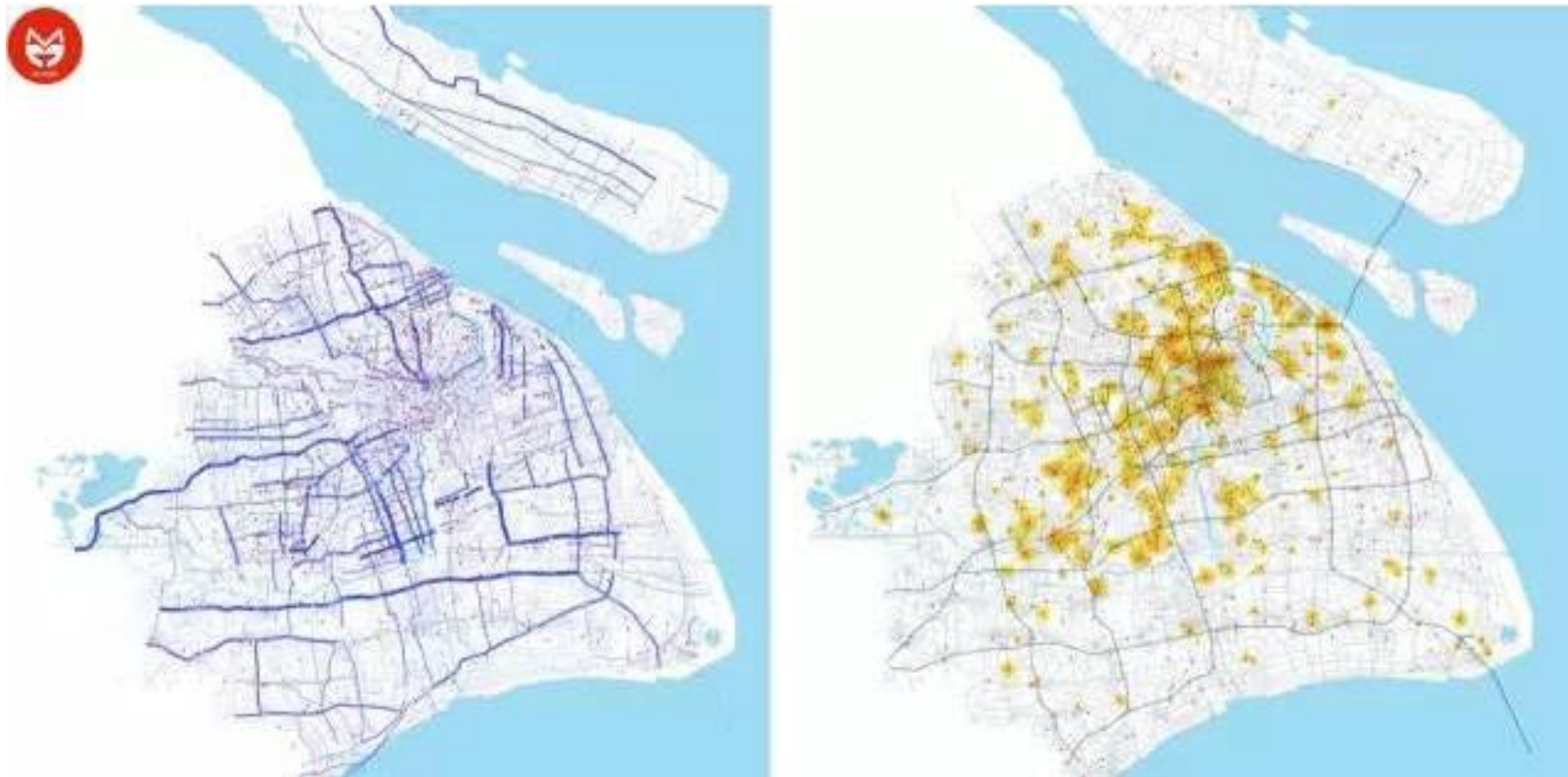
## Customer group study (without labels)



案例：上海市严重交通事故分析

Case: Analysis of serious traffic accidents in Shanghai

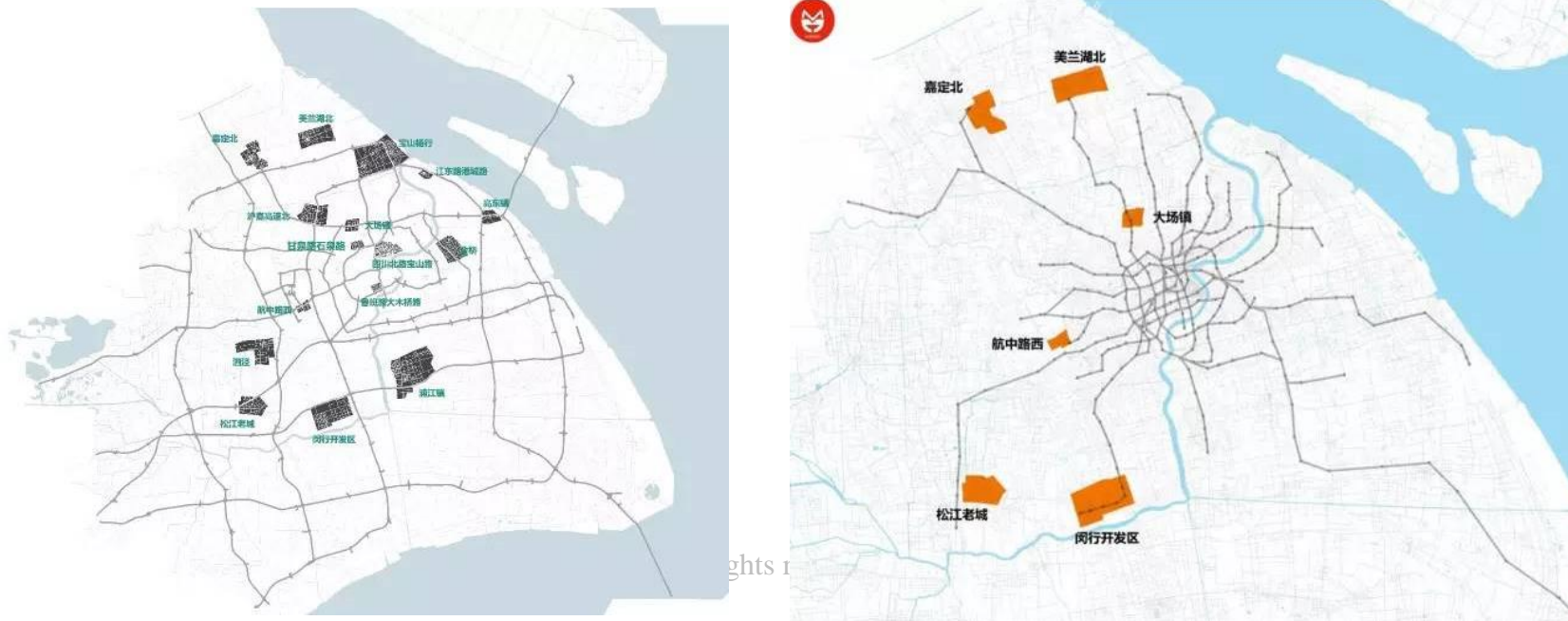
---



案例：上海市严重交通事故分析

Case: Analysis of serious traffic accidents in Shanghai

- There are three main areas prone to traffic accidents in Shanghai:
  - The central urban area with many people and vehicles, large travel volume and complex road conditions;
  - Connecting rail traffic, moped popular suburbs;
  - Along highways and near freight terminals.



# Case: An analysis of various classes in Chinese society

---

- An Analysis of the Various classes of Chinese Society, Volume I, 3-11
  - The landlord class and the comprador class -- the enemy
    - The middle class -- wavering
    - The petty bourgeoisie -- the closest friend
    - The semi-proletariat - the closest friend
    - The proletariat -- the leading force of the revolution
    - Hobo proles – others
- How to analyze the Rural classes, Volume I 127-129





# Types of clustering algorithms

---

- From the perspective of clustering results:
  - **Covering** and non-covering clustering: Each data point belongs to at least one class, which is a covering clustering, otherwise it is a non-covering clustering
  - **Hierarchical** and non-hierarchical clustering: there exist two classes, where one class is a subset of the other and is a hierarchical cluster, otherwise it is a non-hierarchical cluster
  - **Certain** clustering and **fuzzy** clustering: The intersection of any two clusters is empty, and a data point belongs to at most one cluster, which is certain clustering (or hard clustering). Otherwise, if at least one data point belongs to more than one class, it is fuzzy clustering

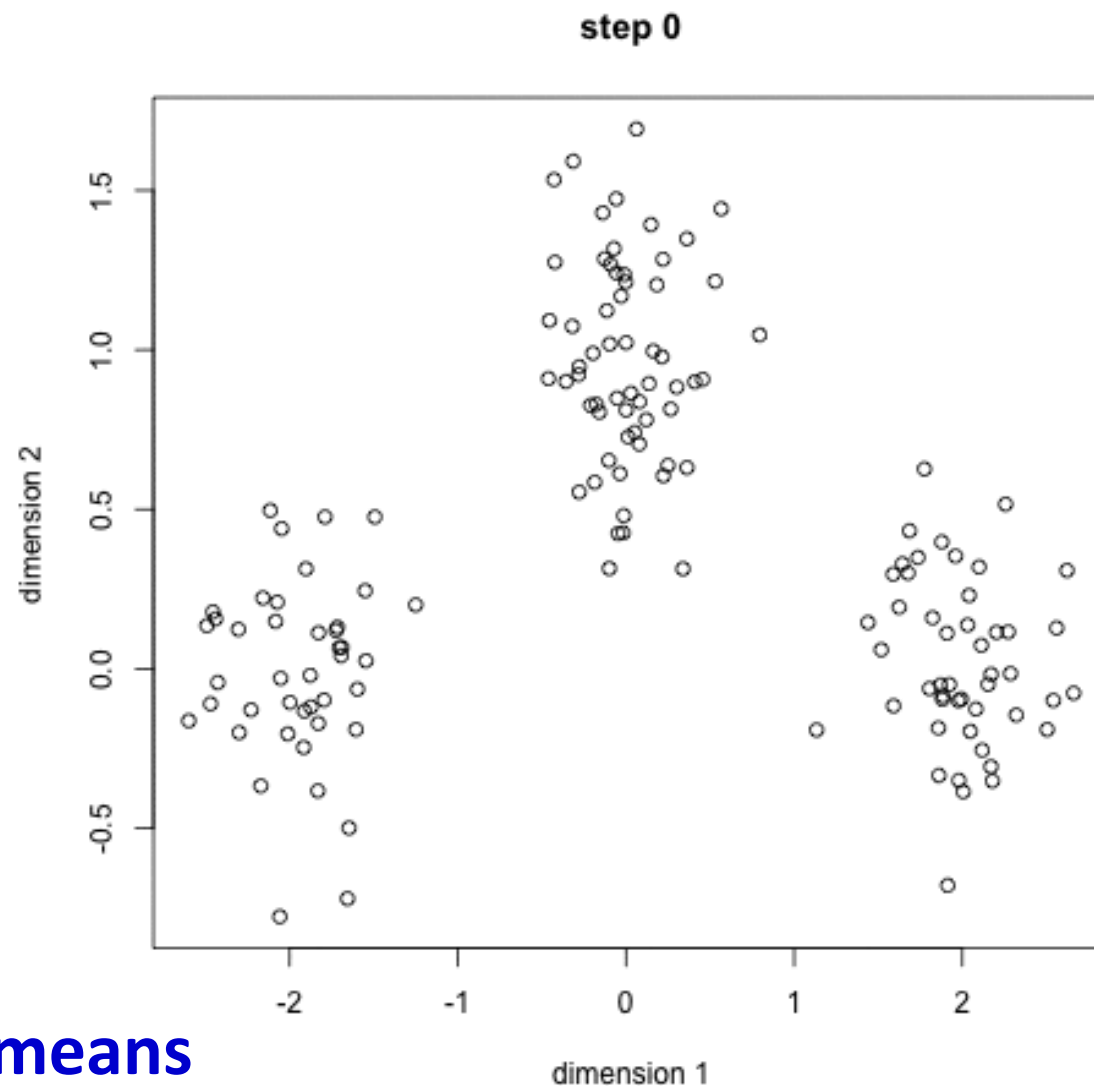
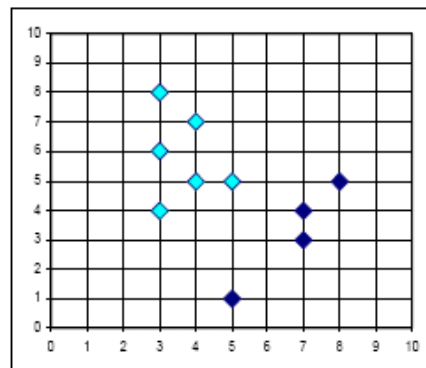
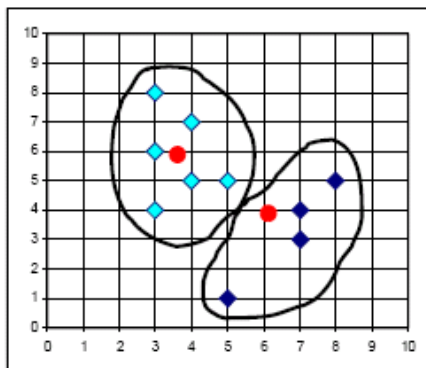
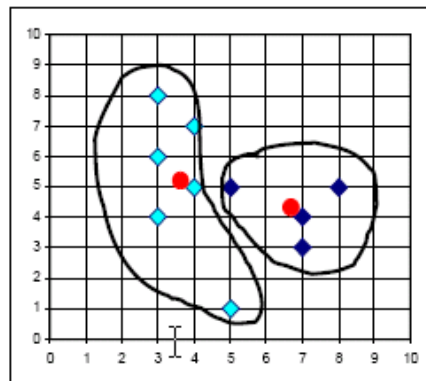
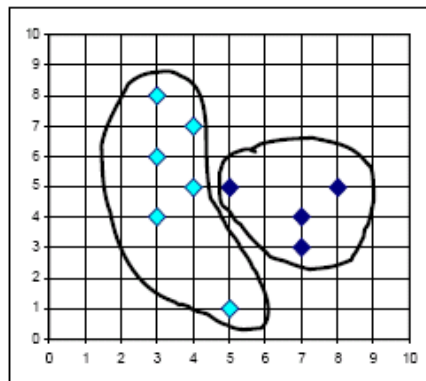
- 
- 从聚类变量类型角度划分Partition from the Angle of clustering variable type
    - 数值型聚类算法（Numerical clustering algorithm）
    - 分类型聚类算法（Categorical clustering algorithm）
    - 混合型聚类算法（Mixed clustering algorithm）
  - 从聚类的原理角度划分Partition from the principle of clustering
    - 划分聚类（Partitional clustering）
    - 层次聚类（Hierarchical clustering）
    - 基于密度的聚类（Density-based clustering）
    - 网格聚类（Grid clustering）



# Clustering algorithm: K-means clustering

---

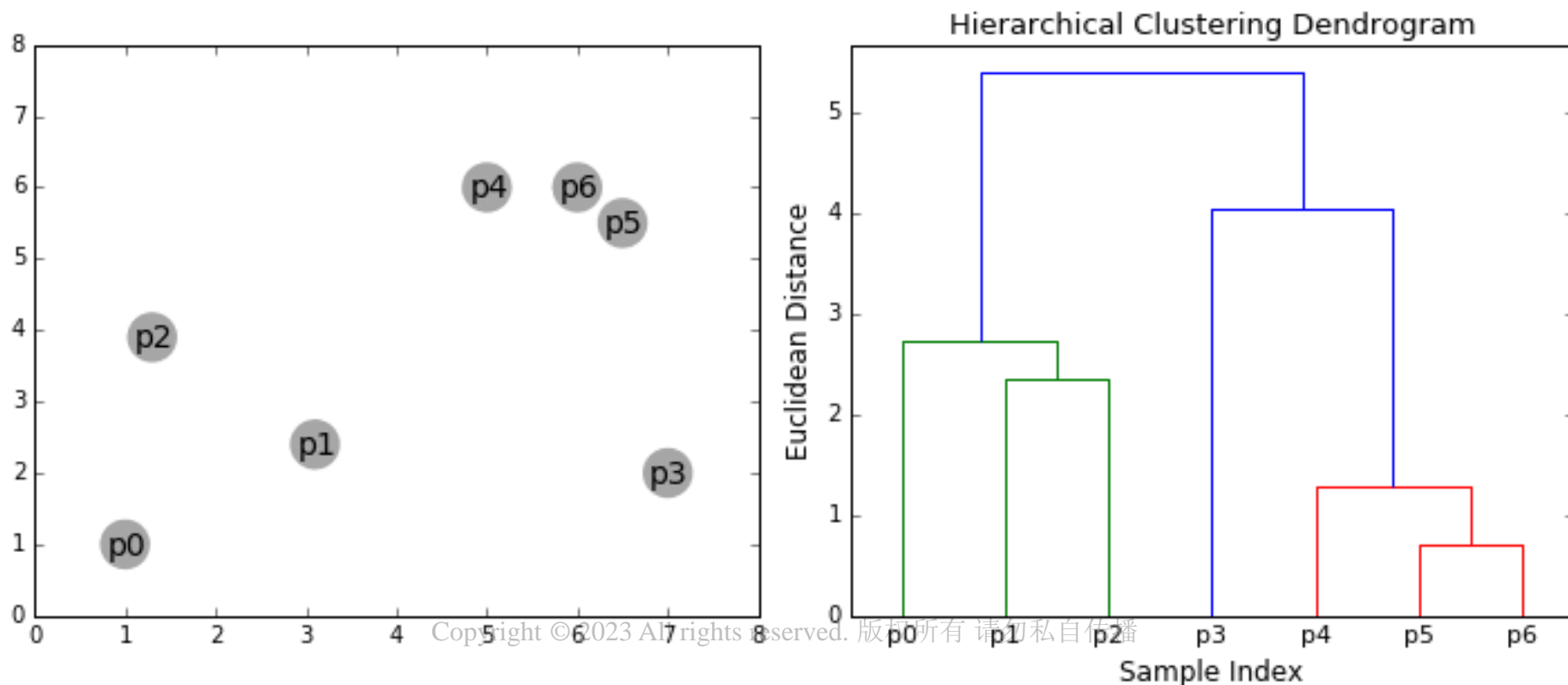
- The k-means algorithm, is one of the most widely used clustering algorithms.
- K-means takes the mean of all data samples in each cluster subset as the representative point of the cluster.
- The main idea of the algorithm is to divide the data set into different categories through the iterative process, so that the criterion function for evaluating the clustering performance is optimal, so that each cluster generated is compact within and independent between clusters.
- K-means algorithm is not suitable for discrete attributes, but it has good clustering effect for continuous attributes



# K-均值聚类示例Example of k-means clustering

# 聚类算法： 层次聚类Hierarchical clustering

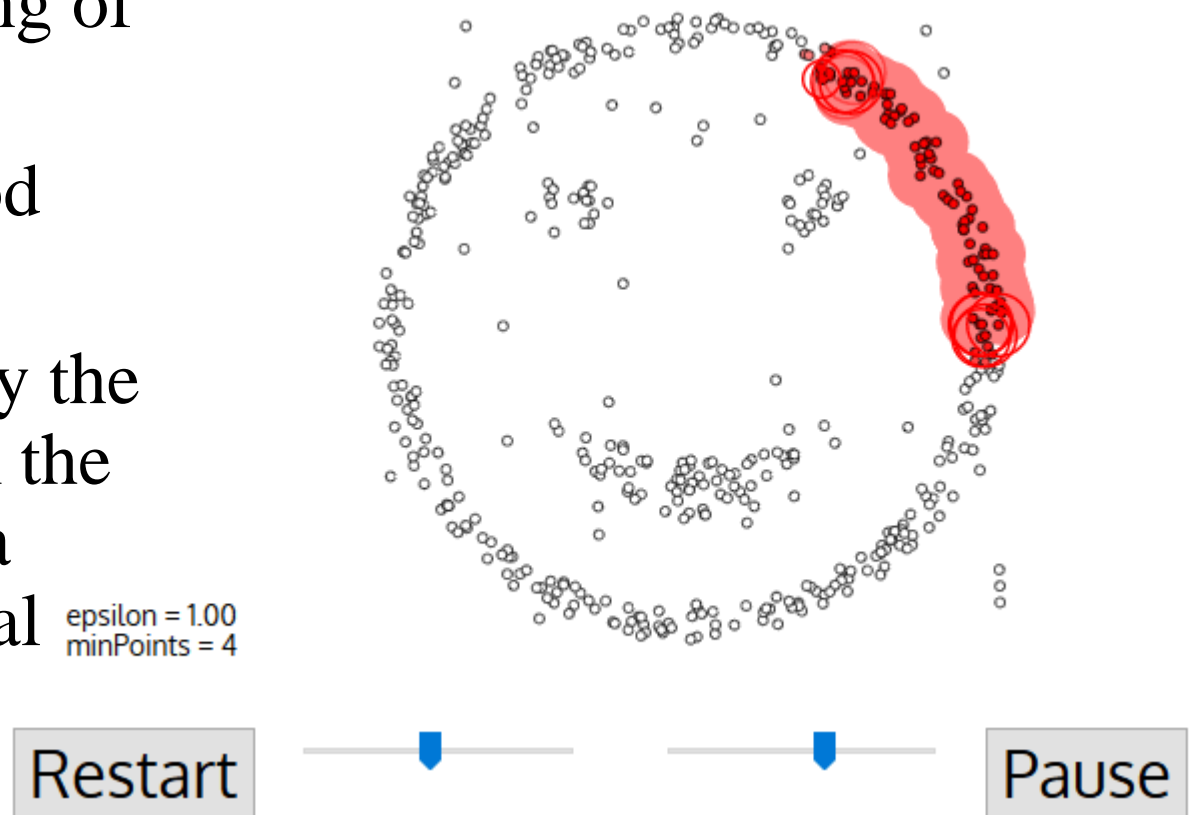
- 逐步合并距离最小的元素/组 Gradually merge the elements/groups with the smallest distance.
- 生成嵌套的类结构， 树图 Generate nested class structures, treemaps.



# Clustering algorithm: DBScan clustering algorithm

---

- Density-Based Spatial Clustering of Applications with Noise
- Density-based clustering method considering noise
- The set of samples connected by the maximum density derived from the density reachability relation is a category, or a cluster, of the final clustering.



## 5. Association rules

---

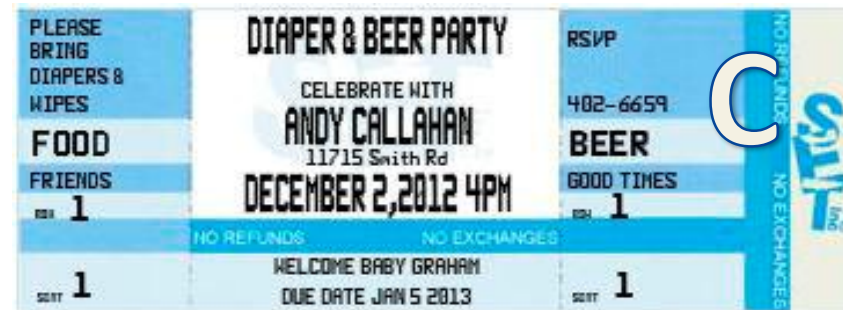
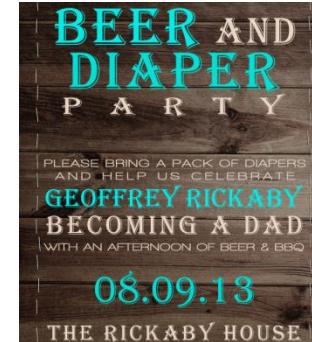
- "Walmart owns the world's largest data warehousing system. To accurately understand customer buying habits in its stores, Walmart conducts shopping basket analysis on its customers to find out which products are frequently purchased together. Walmart's data warehouse contains detailed raw transaction data from its various stores. Using data mining methods on this raw transaction data, Walmart conducts analysis and mining. One unexpected discovery was that 'the most commonly purchased item along with diapers is beer!'"



Case study: Beer & Diaper



- Why “Beer & Diaper”?



## • 不可全信的Not entirely trustworthy都市传奇 (Urban Legend)

- 沃尔玛/一家超市/7-11 Walmart
  - *Sometimes the data can throw up surprises: mining of databases held by 7-Eleven stores in the US revealed a link between purchases of beer and nappies. When they were moved together, sales of both increased, says Williams.*
- 夸大的效果Exaggerated effect
  - *The discount chain moved the beer and snacks such as peanuts and pretzels next to the disposable diapers and increased sales on peanuts and pretzels by more than 27%.*

[啤酒与尿布.pdf 免费高速下载|百度云网盘-分享无限制](#)

文件名: 啤酒与尿布.pdf 文件大小: 61.65M 分享者: 崔占军08 分享时间: 2012-12-19 21:59 下载次数: 428

[pan.baidu.com/share/...](#) 2012-12-19 - 百度快照

[啤酒与尿布 - 搜搜百科](#)

在一家超市中,人们发现了一个特别有趣的现象:尿布和啤酒这两种风马牛不相及的商品居然摆在一起。但这一奇怪的举措居然使尿布和啤酒的销量大幅增加了。这可不...

[baike.soso.com/v3430...](#) 2008-11-01 - 百度快照

[啤酒与尿布\\_滚动新闻\\_新浪财经\\_新浪网](#)

2009年11月27日 - 在一家超市,有个有趣的现象:尿布和啤酒赫然摆在一起出售,但是这个“奇怪的举措”却使尿布和啤酒的销量双双增加了。这是发生在美国沃尔玛连锁店...

[finance.sina.com.cn/r/...](#) 2009-11-27 - 百度快照

[啤酒与尿布\\_iefox\\_新浪博客](#)

在一家超市中,人们发现了一个特别有趣的现象:尿布和啤酒这两种风马牛不相及的商品居然摆在一起。但这一奇怪的举措居然使尿布和啤酒的销量大幅增加了。这可不...

[blog.sina.com.cn/s/bl...](#) 2012-02-21 - 百度快照 - 邀您点评

[啤酒与尿布 - 经管书评 - 人大经济论坛](#)

9条回复 - 发帖时间: 2012年1月30日

在一家超市中,人们发现了一个特别有趣的现象:尿布和啤酒这两种风马牛不相及的商品居然摆在一起。但这一奇怪的举措居然使尿布和啤酒的销量大幅增加了。这可不...

[bbs.pinggu.org/thread...](#) 2012-01-30 - 百度快照

[啤酒与尿布](#)

美国沃尔玛连锁店超市里有个有趣的现象:尿布和啤酒摆在一起出售。这在国人看来或许很难理解,但是这个“奇怪的举措”却使沃尔玛连锁店超市中尿布和啤酒的销量双双...

[www.cicn.com.cn/docro...](#) 2009-12-08 - 百度快照



1

2

3

4

5

6

7

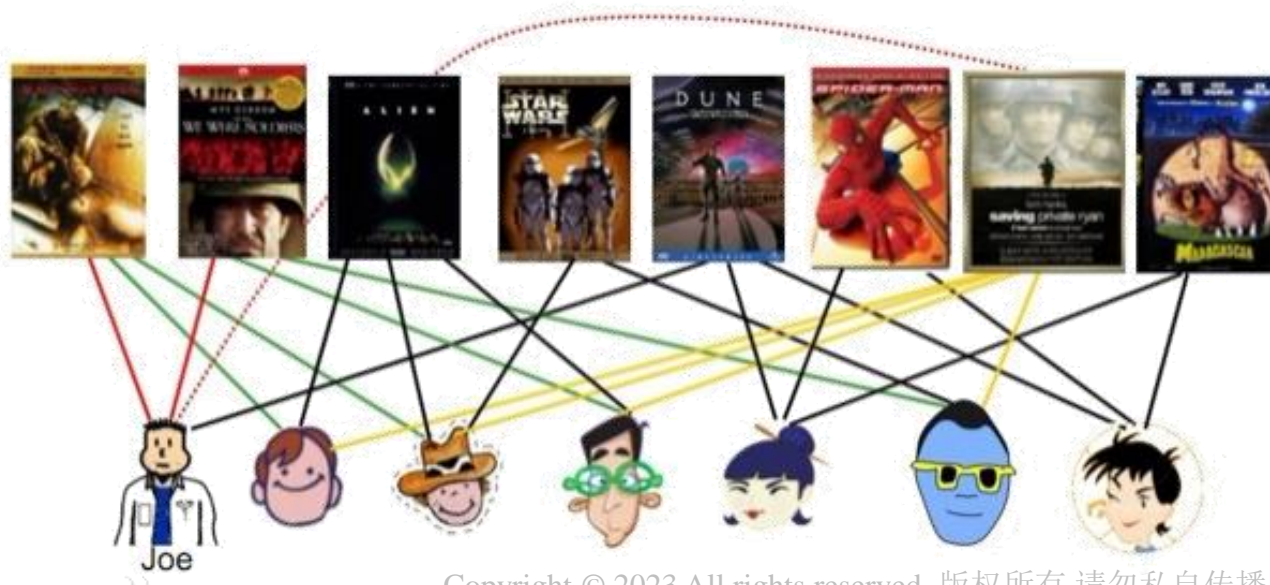
8

9

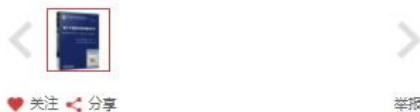
10

下一页>

- 互联网内容推荐Internet content recommendation
  - 为什么刷抖音/快手停不下来Why can't stop?
  - 今日头条/新闻推荐Today's headlines/News recommendations
- 互联网商品推荐Internet product recommendation
  - 关联规则/协同过滤Association rules/Collaborative filtering







企业批量购书

## 基于大数据的商务智能分析

从大数据的角度进行商务智能分析！

[美] 伯特·布瑞吉斯 著，贾岚，殷世嘉，肖青虹，王玲芳 等译

金秋风暴 金秋·无潮不欢

京东价 **¥58.50** [7.5折] [定价 ¥76.00] 降价通知

优惠券 **满300减20**

促销 **加价购** 满12元另加26.90元，或满15元另加16.90元，或满18元另加9.90元，即

可在购物车换购热销商品 详情>>

累计评价 93

增值业务 **助力环保，传递知识，旧书换新**

配 送 至 云南昆明市西山区碧鸡街道 有货

**京东物流** 预约送货 部分收货 送货上门

由 京东 发货，并提供售后服务。11:10前下单，预计明天(09月25日)送达

重 量 0.4kg

服务支持 **放心购** 上门换新 破损包退换 闪电退款

可配送海外49元免基础运费

增值保障 **意外换新** ¥2.50 **2年爱心收** ¥1.00

白条分期 不分期 ¥19.79 × 3期 ¥10.04 × 6期 ¥5.11 × 12期 ¥2.67 × 24期

1 **加入购物车**

温馨提示 支持7天无理由退货

机械工业出版社  
CHINA MACHINE PRESS

木垛图书旗舰店 ¥40.40

博库网旗舰店 ¥40.56

万里路图书专营店 ¥53.80

6个卖家在售

人气单品

七日畅销榜

新书热卖榜



R语言：从数据思维到数据实战

¥70.30



商务智能与分析：决策支持系统（原书第10版）

¥119.10



商务智能（第四版）/清华科技大讲堂

¥49.00



大数据地理信息系统：原理、技术与应用

¥51.60



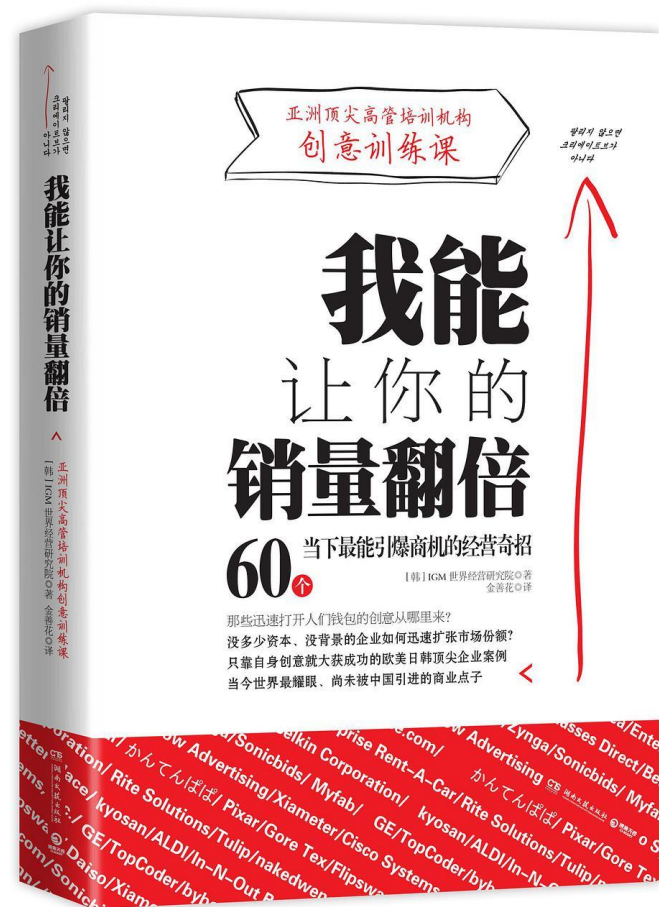
大数据时代（大数据系统研究的先河之作）

¥33.30



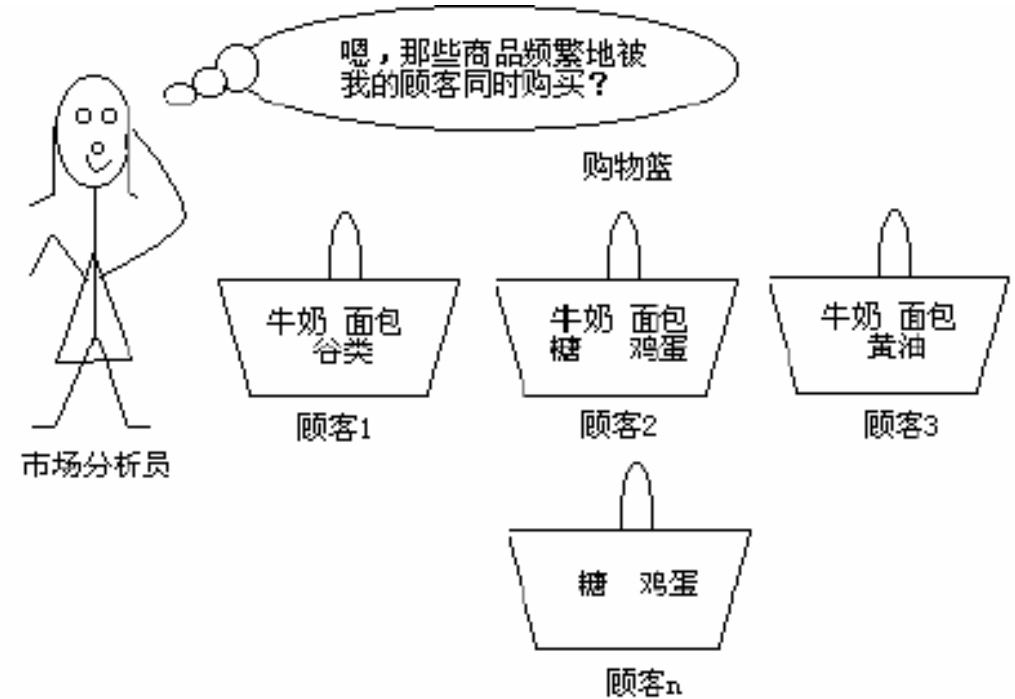
商业数据分析

¥83.70



# Core concepts of simple association rules

- Purpose: To discover patterns of relationships between items and identify their associations.
- Association relationships include: simple association relationships, sequential association relationships.
- The primary technique for association analysis is **Association Rules**.
- It was initially used to study the patterns of products purchased by supermarket customers and is known as market basket analysis.
- An unsupervised learning method.



面包bread→牛奶Milk (S=85%, C=90%)

前项  
Antecedent

后项  
consequent

支持度support

置信度Confidence

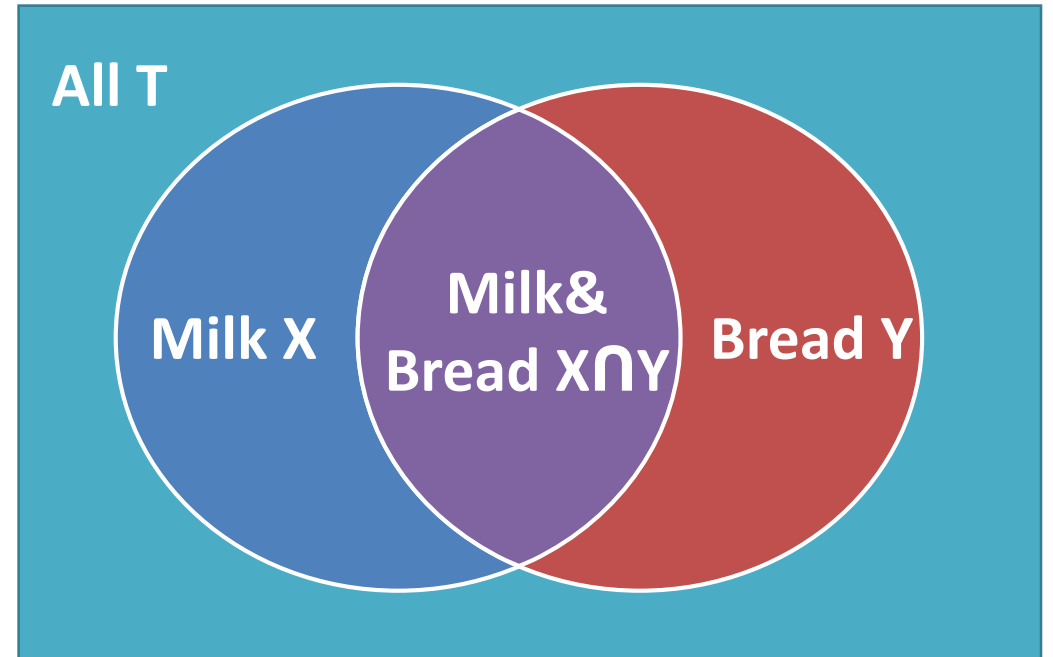
# Antecedents and consequent

---

- General representation of simple association rules:
- $X \rightarrow Y$  (rule support, rule confidence)
- $X$  is the antecedent of a rule, which can be an item or an itemset or a logical expression containing logical and ( $\cap$ ) or ( $\cup$ ) not ( $\neg$ ).
- $Y$  is the successor of a rule, usually an item, indicating some conclusion or fact.
- Examples:
  - Bread  $\rightarrow$  milk
  - Gender (Female)  $\cap$  Income ( $>5000$ )  $\rightarrow$  Brand (A)

# A measure of effectiveness

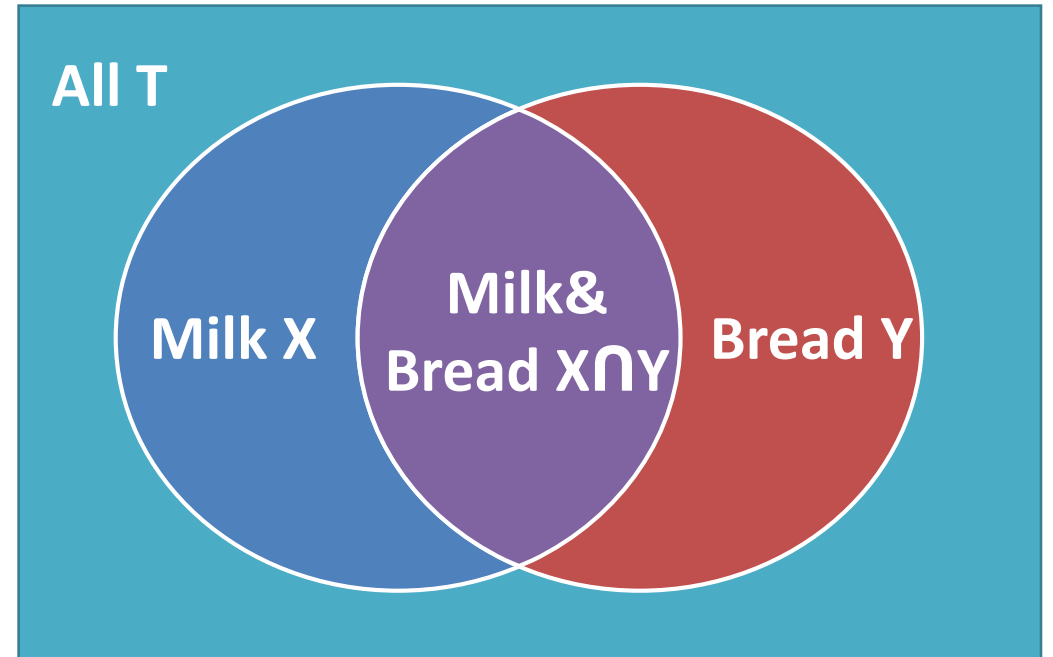
- **Rule Confidence:** A measure of accuracy that describes the probability that a transaction containing item X also contains item Y, reflecting the likelihood of Y given the occurrence of X.
- A high confidence score means that Y is more likely to occur if X occurs.
- Bread  $\rightarrow$  milk (S=85%, C=90%), which means there is a 90% chance that if you buy bread, you will also buy milk.



$$C_{X \rightarrow Y} = \frac{|T(X \cap Y)|}{|T(X)|}$$

Conditional probability

- Rule Support: This measures how common the rule is. Support is the probability that items X and Y occur together.
- Bread  $\rightarrow$  milk (S=85%, C=90%) indicates that there is an 85% probability that a customer will buy both bread and milk.



$$S_{X \rightarrow Y} = \frac{|T(X \cap Y)|}{|T|}$$

# Model Extension 2: Multi-level association rule mining

---

- For many applications, it is difficult to find some strong association rules at the most detailed level of data due to the scattered distribution of data. When we introduce the concept hierarchy, it is possible to mine at a higher level [HF95, SA95]. While the resulting rules at a higher level may be more common information, what is common information for one user may not be common information for another. So data mining should provide such a function of mining at multiple levels.
- Classification of multi-level association rules: According to the levels involved in the rules, multi-level association rules can be divided into same-level association rules and inter-level association rules.
- The framework of "support-confidence" can be used to mine multi-level association rules. However, there are a few things to consider when it comes to support Settings.

## Model Extension 3: Multi-dimensional association rule mining

---

- For multidimensional database, there is a class of multidimensional association rules in addition to intra-dimensional association rules. For example:
  - Age (" 20... 30 ") Occupation (" student ")  $\rightarrow$  Purchase (" laptop ")
  - There are three dimensions: age, occupation, and purchase.
- According to whether the same dimension is allowed to reappear, it can be further subdivided into inter-dimension association rules (dimensions are not allowed to reappear) and mixed-dimension association rules (dimensions are allowed to appear in the left and right of the rule).
  - Age (" 20... 30 ") Buy (" Laptop ")  $\rightarrow$  Buy (" Printer ")



# 4. 机器学习模型的评估及弱点

## Evaluation of Machine Learning Models

# 1. 混淆矩阵（Confusion Matrix）

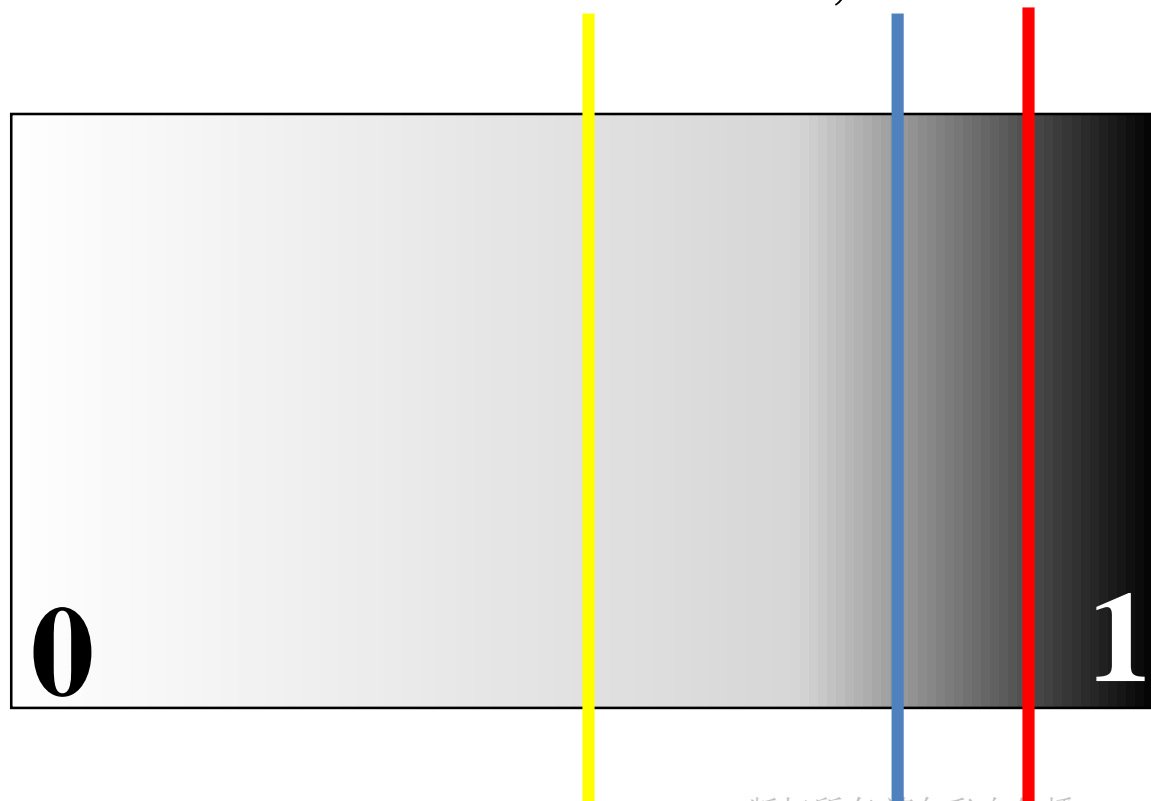
---

混淆矩阵 Confusion Matrix		预测情况Predict		求和Sum
		1 Positive	0 Negative	
真实情况 True	1	TP	FN	TP+FN
	0	FP	TN	FP+TN
求和Sum		P	N	总样本数All

# 分类器的本质 The essence of Classifiers

---

- 按照分数排队，决定阈值，阈值以上为Positive
- Queue by score and decide the threshold, above which is Positive



举例：假设逻辑回归模型输出0.9及以上认为购买

混淆矩阵 Confusion Matrix		预测情况 模型得分 $\geq 0.9$ ?		求和
		分数 $\geq 0.9$ Positive	分数 $< 0.9$ Negative	
真实情况 是否购买	购买	80 (TP)	20 (FN)	100 (TP+FN)
	不购买	400 (FP)	500 (TN)	900 (FP+TN)
求和		480 (P)	520 (N)	1000 (总样本数)

Example: suppose a logistic regression model output of 0.9 and above is considered a purchase

---

Confusion Matrix		Predict Score $\geq$ 0.9?		Sum
		Score $\geq$ 0.9 Positive	Score $<$ 0.9 Negative	
True Whether to buy	Yes	80 (TP)	20 (FN)	100 (TP+FN)
	No	400 (FP)	500 (TN)	900 (FP+TN)
Sum		480 (P)	520 (N)	1000 (All samples)

# 准确率、查准率（Precision）

- 总体正确率(Accuracy)=(TP+TN)/总样本数All=580/1000=0.58
- 预测为1的准确率(Precision)=TP/P=80/480=0.17 (An accuracy of 1 is predicted)

混淆矩阵 Confusion Matrix		预测情况Predict 模型得分 Score>=0.9?		求和Sum
		分数Score>=0.9 Positive	分数Score<0.9 Negative	
真实情况True 是否购买 Whether to buy	购买Yes	80 (TP)	20 (FN)	100 (TP+FN)
	不购买No	400 (FP)	500 (TN)	900 (FP+TN)
求和Sum		480 (P)	520 (N)	1000 (总样本数 All Samples)

# 模型更好还是更差？ Better or worse

- 正确率Accuracy=(TP+TN)/总样本数All=810/1000=0.81
- 准确率Precision=TP/P=10/110=0.09

Accuracy 不重要  
我们真正关心的是准确率

混淆矩阵 Confusion Matrix		预测情况Predict 模型得分 Score>=0.9?		Precision 重要 求和Sum
		分数Score>=0.9 Positive	分数Score<0.9 Negative	
真实情况True 是否购买 Whether to buy	购买Yes	80 (TP)	20 (FN)	100 (TP+FN)
	不购买No	400 (FP)	500 (TN)	900 (FP+TN)
求和Sum		480 (P)	520 (N)	1000 (总样本数 All Samples)



# 召回率、查全率（Recall）

- 召回率(Recall)=TP/(TP+FN)=80/100=0.8
- 准确率(Precision)=TP/P=80/480=0.17

混淆矩阵 Confusion Matrix		预测情况Predict 模型得分 Score>=0.9?		求和Sum
		分数Score>=0.9 Positive	分数Score<0.9 Negative	
真实情况True 是否购买 Whether to buy	购买Yes	80 (TP)	20 (FN)	100 (TP+FN)
	不购买No	400 (FP)	500 (TN)	900 (FP+TN)
求和Sum		480 (P)	520 (N)	1000 (总样本数 All Samples)



# 模型更好还是更差？ Better or worse?

- 召回率(Recall)=TP/(TP+FN)=90/100=0.9 
  - 准确率(Precision)=TP/P=90/690=0.13 
- 召回率 Recall  
和准确率 Precision  
不得两全 此消彼长 can't  
have one or the other

混淆矩阵 Confusion Matrix		预测情况Predict 模型得分 Score>=0.9?		求和Sum
		分数Score>=0.9 Positive	分数Score<0.9 Negative	
真实情况True 是否购买 Whether to buy	购买Yes	80 (TP)	20 (FN)	100 (TP+FN)
	不购买No	400 (FP)	500 (TN)	900 (FP+TN)
求和Sum		480 (P)	520 (N)	1000 (总样本数 All Samples)

# 如何决定模型的阈值？

## How to decide the threshold of the model?

---

- 选择 得分Score>0.9 还是or 得分Score>0.8?

- 理论方法——F1得分 
$$F_1 \text{ Score} = \frac{2 * Precision * Recall}{Precision + Recall}$$

- 例：

- 得分Score>0.9时，  $F1 = 2 * 0.17 * 0.8 / (0.17 + 0.8) = 0.28$

- 得分Score>0.8时，  $F1 = 2 * 0.13 * 0.9 / (0.13 + 0.9) = 0.23$

- **其实没必要Not necessary**

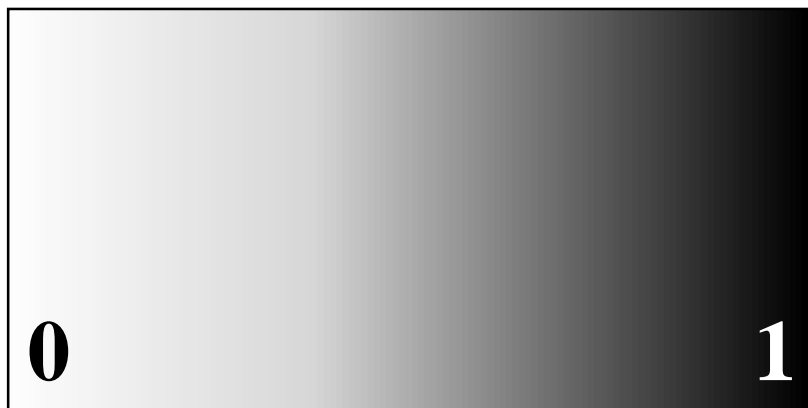
- 
- According to the specific situation, determine the scope of work.
  - The scope is large, the results are many, and the accuracy is low.
  - Small scope, few results, high accuracy.

# 逻辑回归/其它机器学习模型能不能做到100%准确？

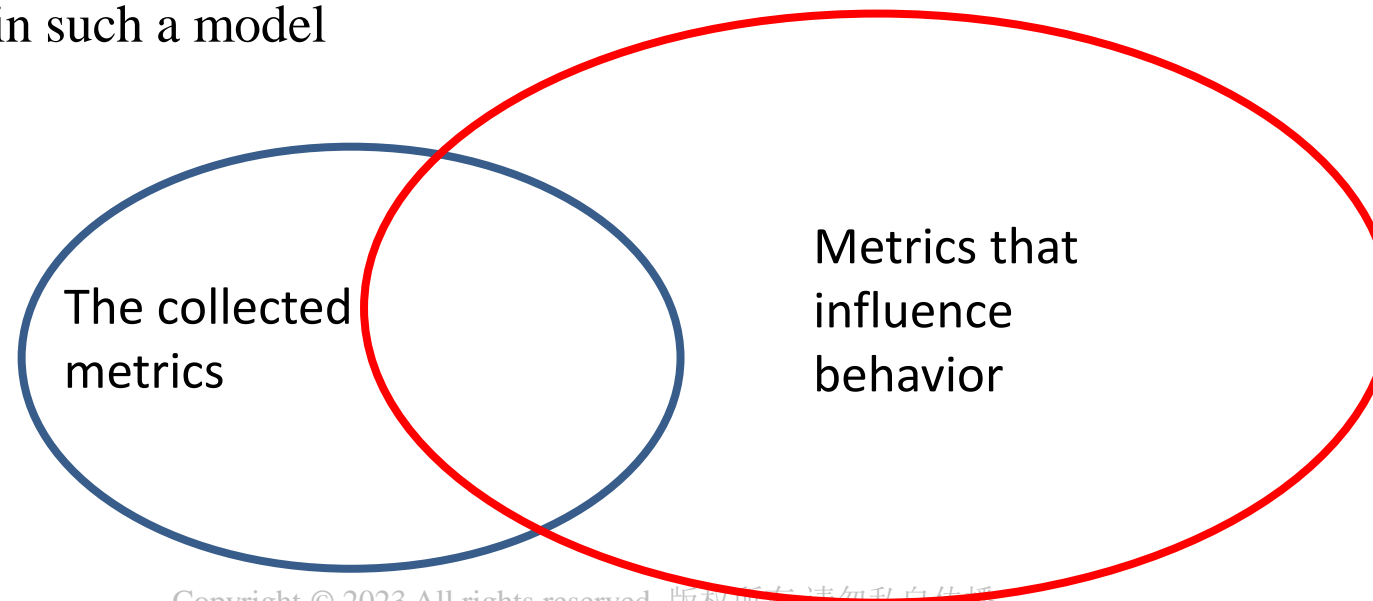
Logistic regression/other machine learning model can achieve 100% accurate?

---

- 有没有一个足够牛模型，可以黑白分明？ Is there a model that is good enough to be black and white?

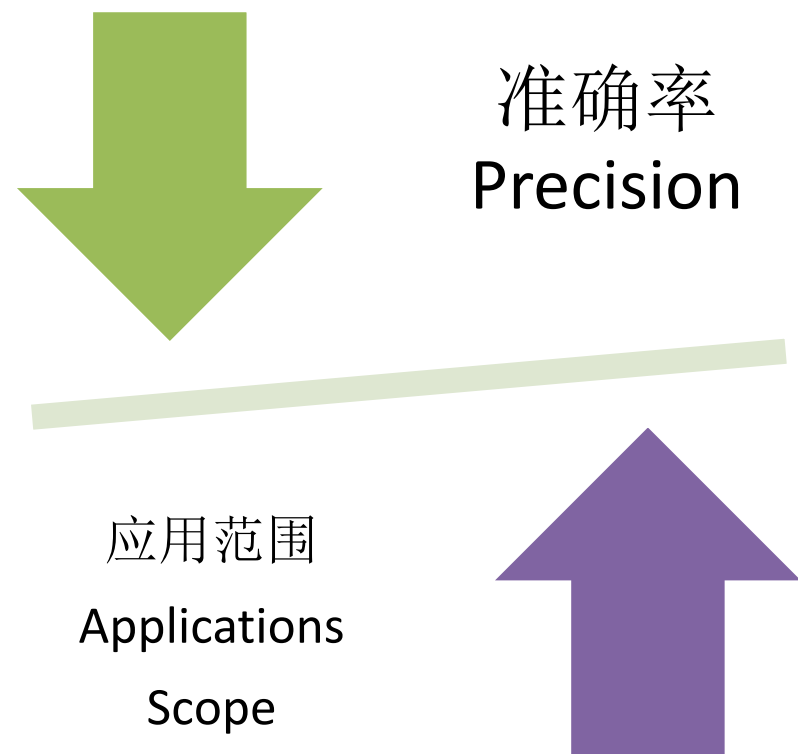
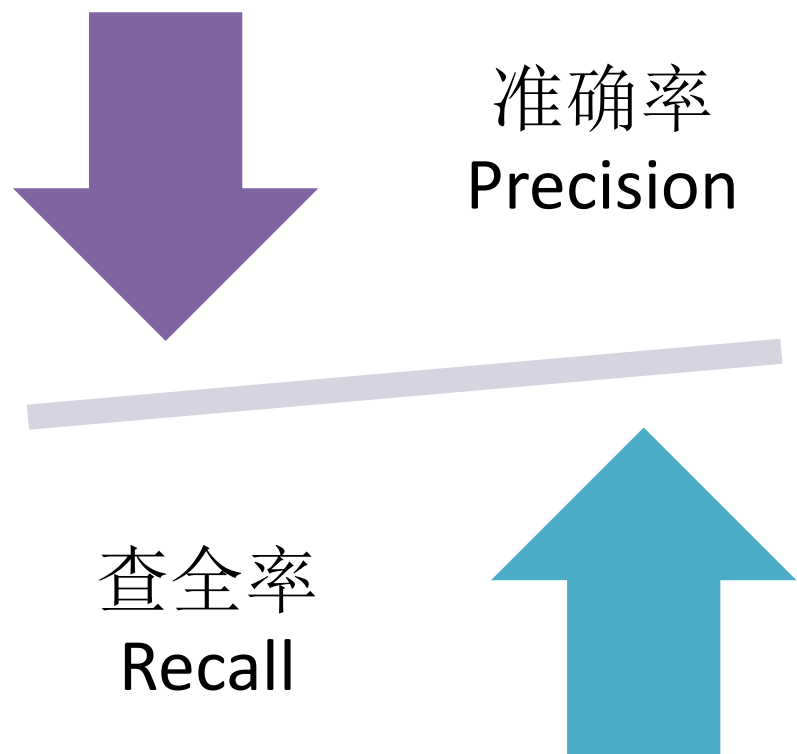


- 
- At the model training level, it's easy!!
    - It is easy to achieve 100% fit between the model and the training data by repeatedly training the model with the collected metrics
  - At the practical level of the model, extremely difficult!!
    - There are so many real metrics that it is impossible to collect them completely, and it is impossible to train such a model



# 模型的权衡Model trade-offs

---



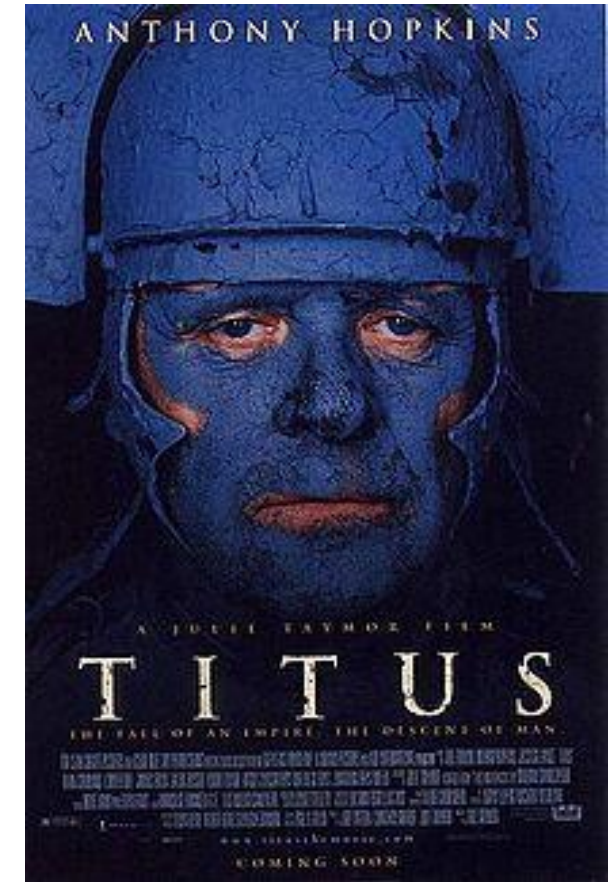
# 5. 机器学习与商业模式的结合

## Machine Learning and Business Models

# 1. Case study: A Differential pricing experiment

~~\$26.24~~ \$22.74

- Streitfeld, David. "On the web, price tags blur: What you pay could depend on who you are." *The Washington Post* September 27 (2000).
- Amazon conducted a dynamic pricing experiment with 68 DVD discs. During the experiment, Amazon determined the pricing levels for these 68 discs based on demographic information of potential customers, their shopping history on Amazon, online behavior, and the software systems they used for internet access.
- For a DVD titled "Titus," the pricing for new customers was set at \$26.24, while for returning customers showing interest in the DVD, the price was \$22.74.
- Through this pricing strategy, some customers paid a higher price than others, allowing Amazon to increase its gross profit margin.





- 
- Less than a month into the implementation of differential pricing, attentive consumers discovered the secret:
    - "An Amazon customer stumbled upon a DVD he wanted to buy for \$26.24, but when he cleaned his computer and returned to Amazon, the same DVD was priced at \$22.74."
  - Hundreds of DVD consumers got the word out through a community of music lovers called DVDTalk ([www.dvdtalk.com](http://www.dvdtalk.com))
    - Of course, the customers who pay the high prices are complaining, and they have taken to the Internet to criticize Amazon's practices in fierce words.
    - Some people openly say they will never buy anything from Amazon in the future.



- 
- Amazon only recently revealed that it tracks and records consumers' shopping habits and behavior on its site.
  - After the incident came to light, consumers and the media began to wonder whether Amazon was using the data it collected on consumers as the basis for its price adjustments, and such speculation connected Amazon's price incident with sensitive online privacy issues.



- 
- Bezos, Amazon's chief executive, had to take the crisis into his own hand, pointing out that Amazon's price changes are random, regardless of who the customer is, and that the purpose of the price experiment is only to test the customer's response to different discounts. Amazon "does not use demographic data to make dynamic pricing, neither now nor in the future."
  - Bezos publicly apologized to customers for any distress caused by the incident. Not only that, but Amazon has also tried to save people's minds with real action. Amazon has promised to give the largest discount to all consumers who bought the 68 DVDS during the price test period. According to incomplete statistics, at least 6,896 customers who did not buy DVDS at the lowest discount price have been repaid by Amazon.



## 2. 案例：Farecast预测机票票价

### Case study: Farecast predicts airline ticket prices

---

- Etzioni, O., R. Tuchinda, C. A. Knoblock and A. Yates (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM.
- 维克托·迈尔-舍恩伯格 《大数据时代：生活、工作与思维的大变革》



Oren Etzioni  
University of Washington  
Computer Science & Engineering  
Director of Turing Center





- 
- "In 2003, Oren Etzioni was getting ready to fly from Seattle to Los Angeles to attend her brother's wedding. He knew the earlier he booked the cheaper the flight, so he booked a flight to Los Angeles online months before the big day."



- "On the plane, Etzioni curiously asked the passenger next to him how much he had paid for his ticket. He was very angry when he learned that even though the man had bought his ticket later, the ticket was much cheaper than his own. He asked a few other passengers and found that they all had cheaper tickets than he did."



- 
- Etzioni, O., R. Tuchinda, C. A. Knoblock and A. Yates (2003). To buy or not to buy: mining airfare data to minimize ticket purchase price. Proceedings of the ninth **ACM SIGKDD** international conference on Knowledge discovery and data mining, ACM.



This paper reports on a pilot study in the domain of airline ticket prices where we recorded over 12,000 price observations over a 41 day period. When trained on this data, Hamlet — our multi-strategy data mining algorithm — generated a predictive model that saved 607 simulated passengers \$283,904 by advising them when to buy and when to postpone ticket purchases. Remarkably, a clairvoyant algorithm with complete knowledge of future prices could save at most \$320,572 in our simulation, thus HAMLET's savings were 88.6% of optimal. The algorithm's savings of \$283,904 represents an average savings of 27.1% per simulated passenger for whom savings are possible. Our pilot study suggests that mining of price data available over the web has the potential to save consumers substantial sums of money per annum, at least until corporations begin to fight back.

- 
- "Etzioni created a prediction system that helped virtual passengers save a lot of money. The system is based on a sample of 12,000 prices over 41 days, scraped from a travel website."
  - "The prediction system doesn't say why, it can only speculate about what will happen. That is, it does not know what factors contribute to the fluctuations in ticket prices. It doesn't know if the price is down because of unsold seats, seasonal reasons, or "not going out on Saturday nights." The system only knows how to use data from other flights to predict future price movements."



- "That little project grew into a venture-capital funded tech startup called Farecast. "By predicting where airfares will go and how much they will rise or fall, Farecast's fare prediction tool helps consumers make the most of their purchases in a way that no other site has made possible before."

The screenshot displays the Farecast website with the following sections:

- Search Fares Nationwide:** Includes input fields for 'From' and 'To', date pickers for 'Leave' (03/23/2007) and 'Return' (03/30/2007), an 'Adults' dropdown (set to 1), and a 'GO' button. A 'Save \$40' badge indicates savings on average with fare predictions. A link 'Vote to add predictions for your city' is also present.
- Know When to Buy – Airfare Predictions:** Features five directional arrows (red up, orange up-right, blue right, green down-right, green down) with corresponding buttons: 'Buy Now. Fares will rise.' and 'Wait. Fares will drop.'
- Know Where to Buy – Airline Websites:** A section with a 'Learn More' link.
- Know When to Travel – Graph View:** Shows a line graph of fare fluctuations over 'Departure Dates' with a price range from \$300 to \$400. It includes 'From' and 'To' dropdown menus and a 'GO' button.
- Introducing Fare Guard – Protect Your Fare:** Promotes a service for \$9.95 to protect the lowest fare, with a 'Learn More' link.



- 技术的接受度、实用性 Technology acceptance, usability

**Refine Results** [ [Reset](#) ]

From:  
Seattle, WA - SEA

To:

Honolulu, HI - HNL ☒

Maui, HI - OGG ☒

Kona, HI - KOA ☒

Kauai, HI - LIH ☒

Select a city

[Vote to add cities](#)

Leave between:

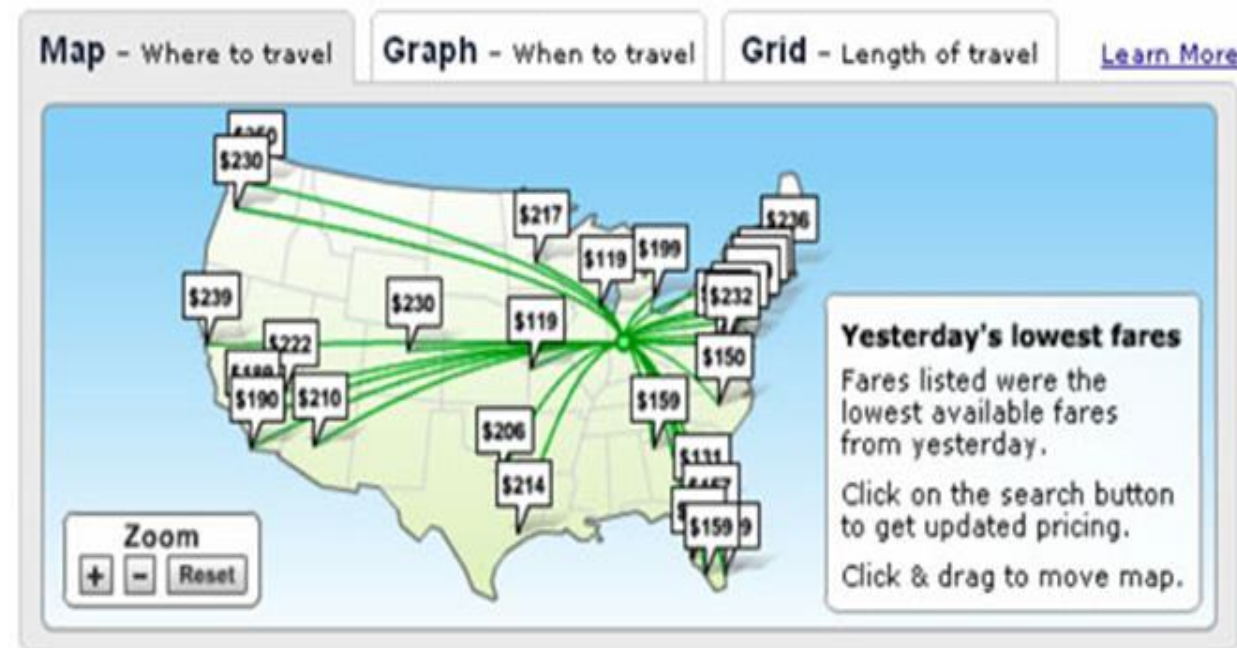
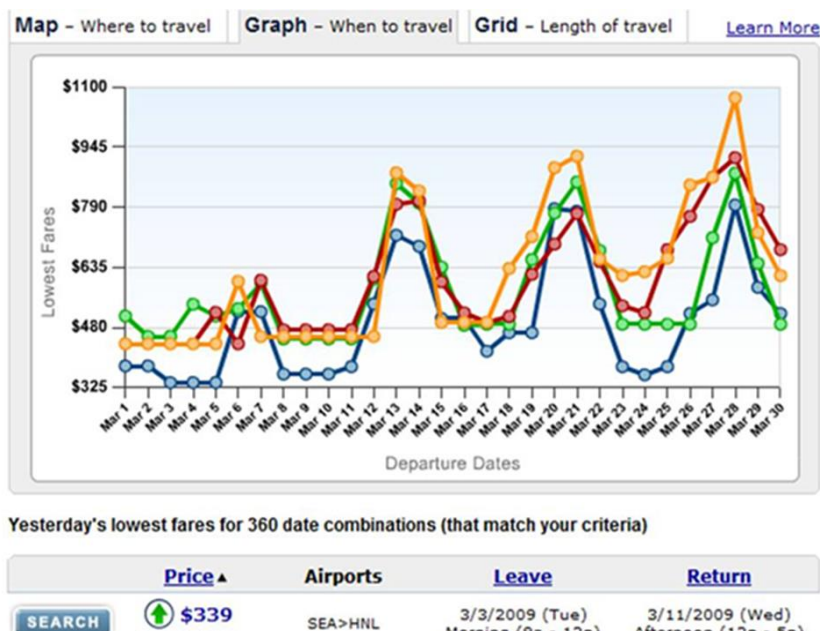
03/01/2009  
and  
3/30/2009

Compare a 30-day range

Trip Length (nights):

2 4 6 8 10 12 14

[ [Reset](#) ]





**Buy Now.**  
Fares will rise.



**Wait.**  
Fares will drop.

# 大数据的“阿基里斯之踵”：准确率

## The Achilles heel of Big Data: accuracy

### 7-Day Low Fare Prediction



**Tip: Wait**

Fares Dropping or Steady.

Confidence: 63%

[Consider Risk](#)

**Catch Fare Drop**

### Daily Low Fare History



### Fare Prediction



**Lowest fares rising \$50+**  
on average over the next 7 days

Confidence: 76%

**Tip: Buy Now.** [Learn More](#)

谢谢！  
Thank you for your attention.

[liuyuewen@xjtu.edu.cn](mailto:liuyuewen@xjtu.edu.cn)

