02 - A gentle introduction to probability and statistics

Hu Chuan-Peng

2023-03-01

Review

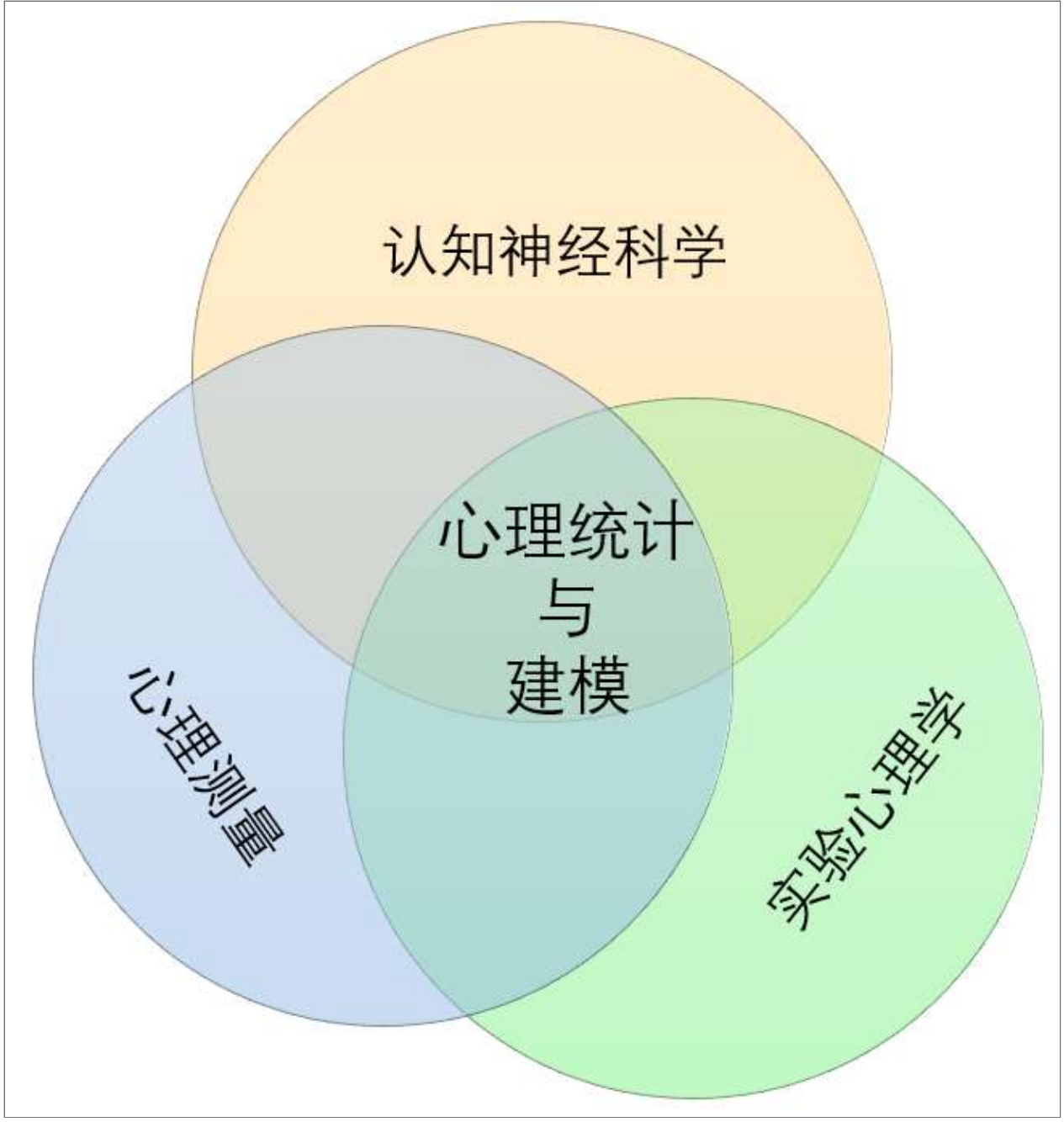
心理统计学的两种视角

统计学

- 在心理学中的应用
- 在经济学中的应用
- 在生物学中的应用
- ...

科学心理学(研究方法)

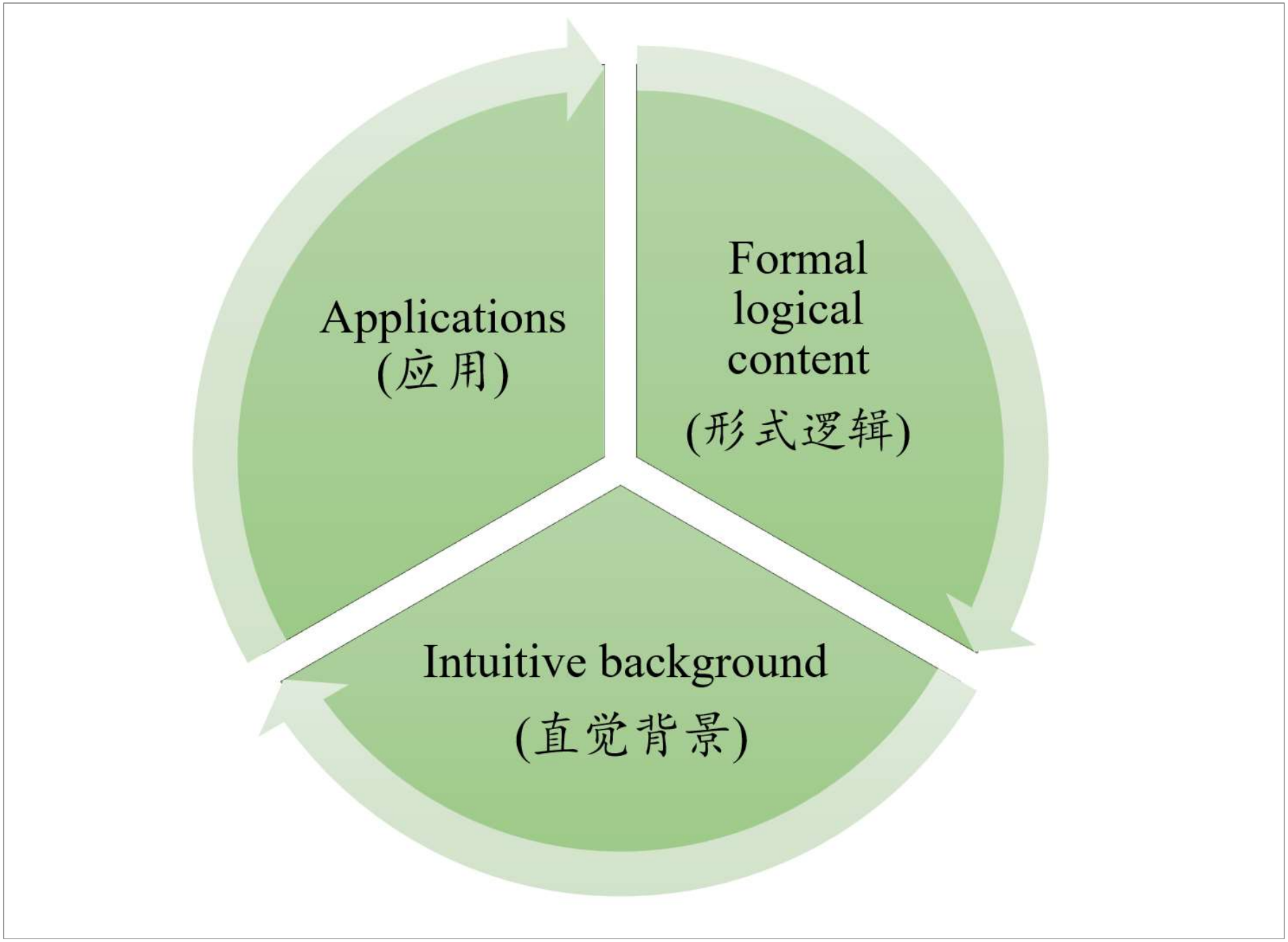
- 心理统计学
- 实验心理学/实验设计
- 心理测量
- 神经成像方法
- ...


认知神经科学
心理统计
与
建模
心理测量
实验心理学

1. THE BACKGROUND

Probability is a mathematical discipline with aims akin to those, for example, of geometry or analytical mechanics. In each field we must carefully distinguish three aspects of the theory: (*a*) the formal logical content, (*b*) the intuitive background, (*c*) the applications. The character, and the charm, of the whole structure cannot be appreciated without considering all three aspects in their proper relation.

(William Feller (1968). An introduction to probability theory and its applications. Chapter 1)



- Application:

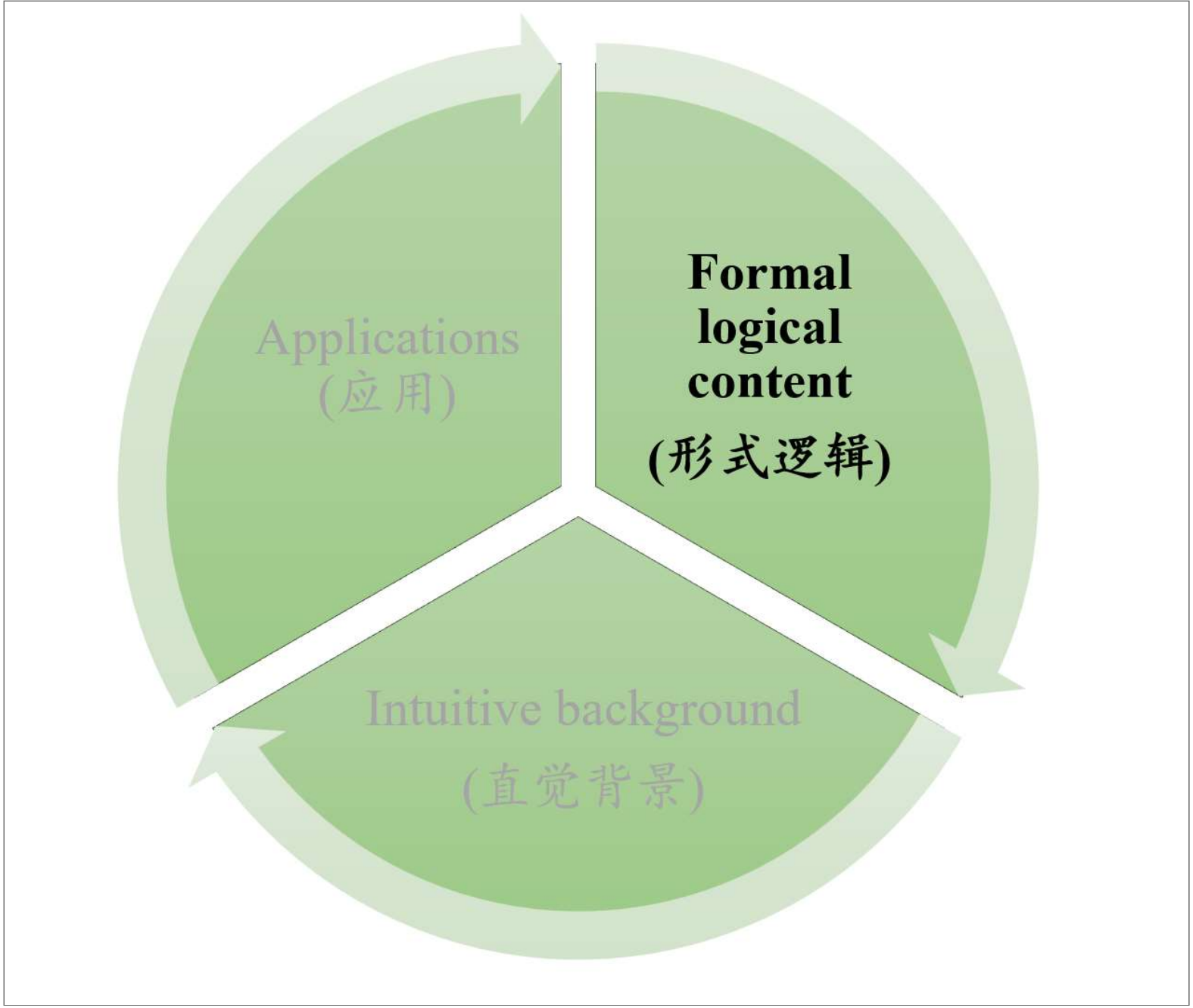
- Different models can describe the same empirical situation.

The manner in which mathematical theories are applied **does not depend on** *preconceived ideas; it is a purposeful technique depending on, and changing with,* **experience.**

日常生活中的现象

- 在选举期间，政治调查机构会进行各种调查，以了解选民的意见和偏好。
- 在医学研究中，进行临床试验检验一个新药是否比现有药物更有效。
- 在股票市场中，投资者可以使用股票价格图来了解市场趋势和预测股票价格的变化。
- 在疫苗试验中，研究人员可能会随机选择一些参与者，以接受实验药物或安慰剂。
- 人们在经济决策中更喜欢短期回报还是长期回报?
- 宇宙中的天体运行有迹可循吗?

概率(Probablity)


Applications
(应用)
Formal
logical
content
(形式逻辑)
Intuitive background
(直觉背景)

当你掷硬币的时候，哪面朝上?

数字(正面) 或 花(反面)

随机现象(Random phenomena)

- 在一定条件下，出现的可能结果不止一个，事前**不能**确切知道哪一个结果一定会出现，但**大量重复试验(trial)**中其结果又具有统计规律的现象;
- **偶然性(Uncertainty)**: 在每一次试验(trial)之前，结果都无法事前确定;
- **规律性(Regularity)**: 相同条件下进行多次重复试验(trial)，实验(experiment)的结果会呈现出统计规律.

> 多次抛硬币，每次的结果是正面朝上还是反面朝上是未知的。但是进行10,000次，大约有一半的次数是正面朝上。

- 在新冠疫情中，感染新冠病毒的情况就是一个典型的**随机现象**。
  - 在同样的条件下，不同的人可能会表现出不同的感染情况。例如，某些人可能会因为免疫系统较强而不易感染，而另一些人则可能因为免疫系统较弱而更容易感染。**（偶然性）**
  - 同一片地区，不同的人在相同的条件下感染了新冠病毒，他们的症状相似。**（规律性）**

开始抛10次硬币，你可以列出实验(experiment)的所有结果吗？

记录下每次抛硬币的结果。

> **正面：数字，反面：花**

样本空间(Sample space, Ω)

- 随机实验(experiment)的所有可能结果构成的集合(set);
- $p(\Omega)$ = 1;
- E.g.1. 当你掷一枚硬币，Ω={正面，反面}
- E.g.2.在新冠疫情的情境中，我们可以将样本空间定义为所有可能感染新冠病毒的人群集合，包括已经感染、未感染但易感染等人群。

你可以预测实验(experiment)中的下一个试次(trial)的结果吗?

事件(Event)

- 实验(experiment)结果的集合(样本空间的一个子集(subset));
  - 一个事件A表示硬币落在正面，用A = {正面} 来表示。
  - 在新冠疫情中，一个事件可以是指所有感染新冠病毒的人群，或是所有在某一时间点内接种了疫苗的人群。

基本事件(Elementary event)

- 在样本空间中只包含一个结果的事件;
- 任何两个基本事件是互斥的;
  - 硬币落在正面或反面的情况，可以用{正面}和{反面}来表示。
  - 某个人是否感染了新冠病毒，或是某个人是否在某一时间点内接种了疫苗。

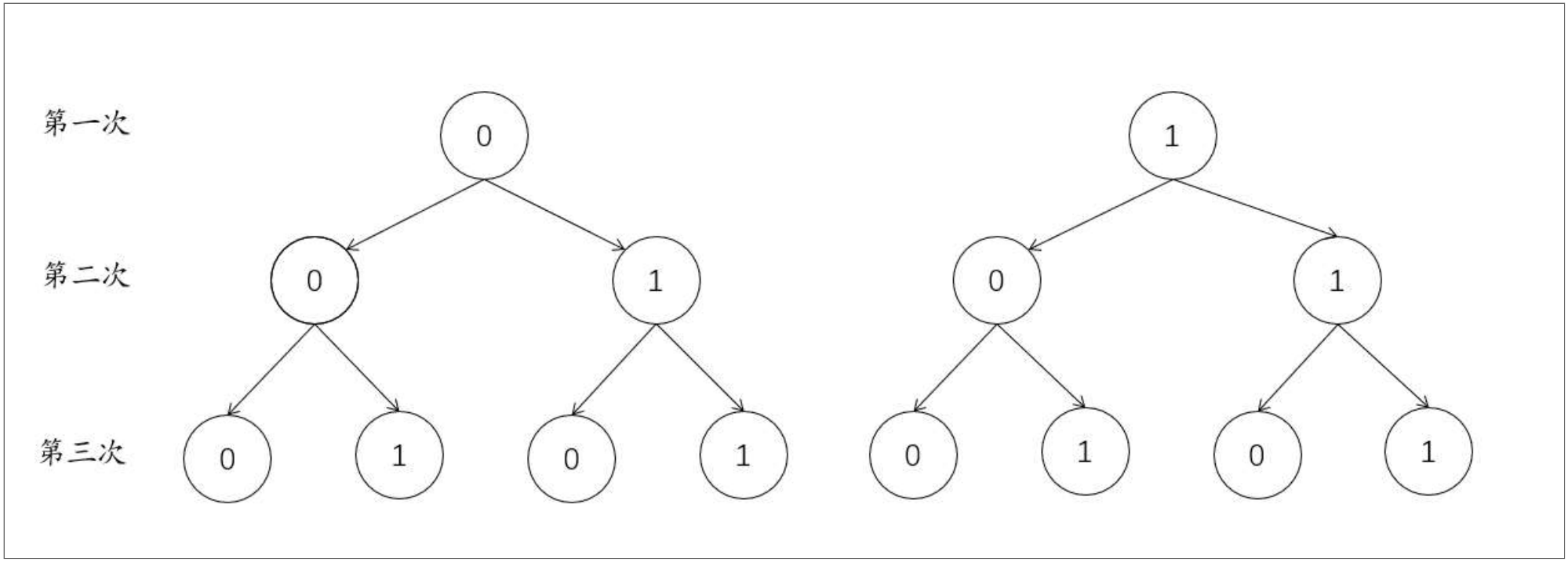
复合事件(Compound event)

- 在样本空间中包含一个以上结果的事件;
- 任何复合事件都可以表示为若干个基本事件的和。
  - 在某一时间点内既接种了疫苗又感染了新冠病毒的人群。
  - 如果要求抛硬币两次，每次都落在正面的概率，则可以定义一个复合事件B，表示两次抛硬币的结果都是正面。用B = {正面,正面} 来表示。

当你掷硬币一次，基本事件是什么呢?

当你掷硬币两次，基本事件是什么呢?

当你掷硬币三次，基本事件是什么呢?



(注："0"表示反面，"1"表示正面)

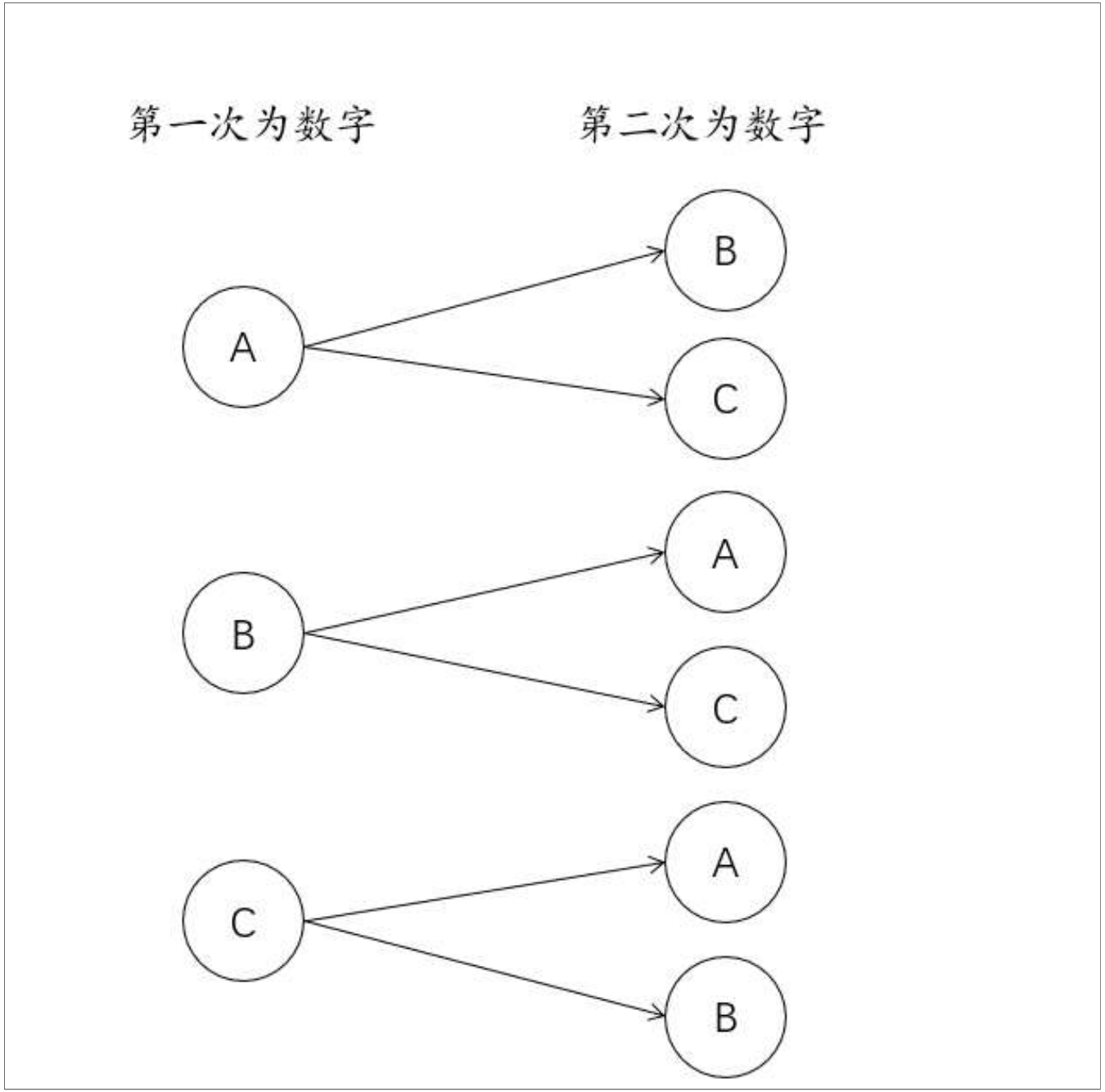
$$2 * 2 * 2 = 8$$

排列(Permutation)

若我们将三枚硬币分别编码为硬币A、B、C

前两次掷硬币结果为*正面*有几种情况?

排列(Permutation)



$$A_3^2 = 3 * 2 = 6$$

排列(Permutation)

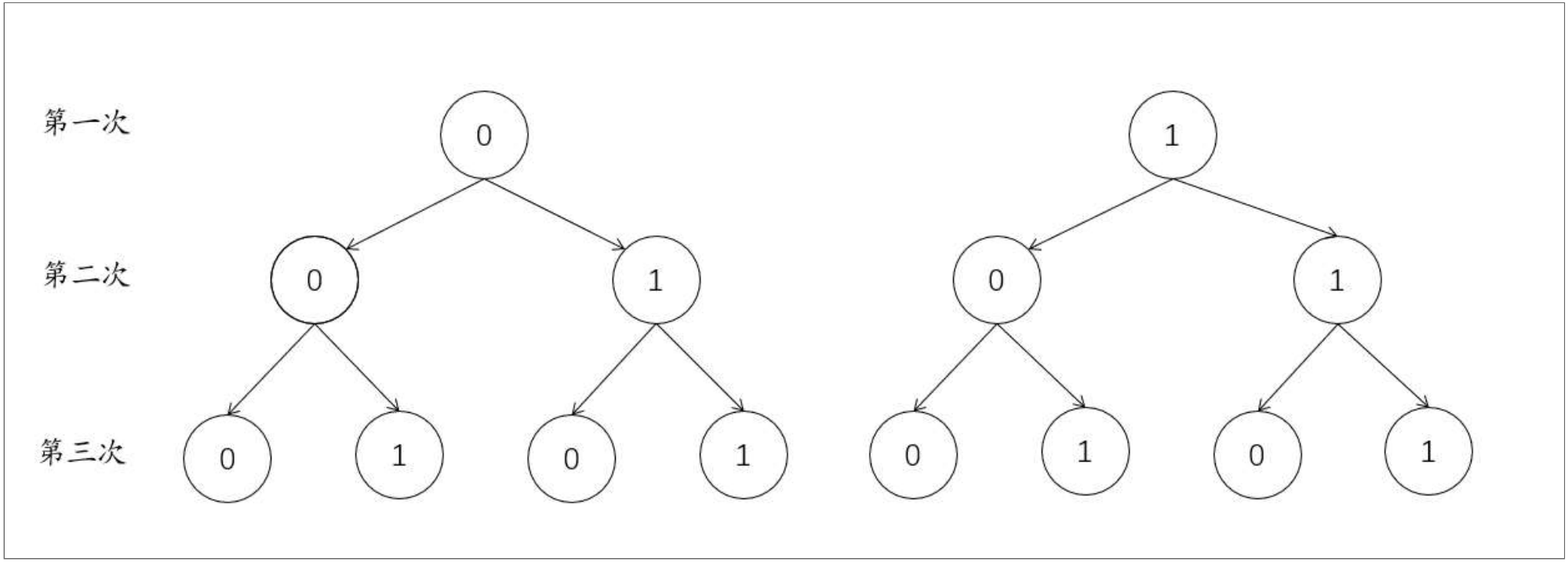
- 从给定个数的元素中取出指定个数的元素进行排序;
- 从n个不同元素中，任取m(m≤n,m与n均为自然数,下同）个不同的元素按照一定的顺序排成一列，叫做从n个不同元素中取出m个元素的一个排列;
- 从n个不同元素中取出m(m≤n）个元素的所有排列的个数，叫做从n个不同元素中取出m个元素的排列数，用符号 A_n^m 表示。

$$A_n^m = n!/(n-m)! = n(n-1)(n-2)\dots(n-m+1)$$

组合(Combination)

掷硬币三次，2次正面朝上的可能有几种?

组合(Combination)



(注："0"表示反面，"1"表示正面)

3

组合(Combination)

- 从给定个数的元素中仅仅取出指定个数的元素，不考虑排序;
- 从n个不同元素中，任取m(m≤n）个元素并成一组，叫做从n个不同元素中取出m个元素的一个组合;
- 从n个不同元素中取出m(m≤n）个元素的所有组合的个数，叫做从n个不同元素中取出m个元素的组合数。用符号 C_n^m 表示。

$$C_n^m = n!/(n-m)!m! = n(n-1)(n-2)\dots(n-m+1)/m(m-1)(m-2)\dots 1$$

我们抛出的硬币公平吗？也就是说，抛出硬币时，数字朝上的可能性是多少？

- 让我们掷硬币20次
- 让我们无限次掷硬币

频率(Frequency)&概率(Probablity)

频率(Frequency)

- 实验过程中事件发生的次数，除以实验的总次数，一般记为f。

概率(Probablity)

- 对一个事件发生的可能性有多大，或一个命题是真的可能性有多大的数字描述，一般记为p。

频率(Frequency)&概率(Probablity)

- 假定我们掷硬币n次，正面朝上的次数为n_1次，则正面朝上的频率为：$f(\text{数字}) = \frac{n_1}{n}$;
- 当我们掷硬币的行为重复无数次，频率就接近于概率：
  $\lim_{n \to +\infty} f(\text{数字}) = \frac{n_1}{n} = p(\text{数字})$

频率(Frequency)&概率(Probablity)

许多数学家都进行过这项实验：

	n	n_1	f
德摩根(De Morgan)	2048	1061	0.5181
蒲丰(Buffon)	4040	2048	0.5069
皮尔逊(Karl Pearson)	24000	12012	0.5005

频率接近于0.5

概率的性质

- $p(\Omega) = 1$
- $0 \leq p(A) \leq 1$
- $p(A + B) = p(A) + p(B) - p(A \cap B)$

独立事件(Independent event)

- 如果两个事件A和B是独立的

$$p(A\,and\,B) = p(A \cap B) = p(A)p(B)$$

$$p(A\,or\,B) = p(A \cup B) = p(A) + p(B) - p(A \cap B)$$

互不相容事件(Mutually exclusive event)

- 如果两个事件A和B是互不相容的

$$p(A\,and\,B) = p(A \cap B) = 0$$

$$p(A\,or\,B) = p(A \cup B) = p(A) + p(B) - p(A \cap B)$$
$$= p(A) + p(B) - 0 = p(A) + p(B)$$

概率的加法法则(Addition rules in probability)

- 若A和B为独立事件

$$p(A) + p(B) = p(A \cup B) - p(A \cap B)$$

- 若A和B为互不相容事件

$$p(A) + p(B) = p(A \cup B)$$

概率的乘法法则(Multiplication rules in probability)

- 若A和B为独立事件，A和B同时发生的概率为

$$p(A \cap B) = p(A)p(B)$$

- 若A和B彼此不独立，A和B同时发生的概率为

$$p(A \cup B) = p(A)p(B|A)$$

> p(B|A) 意为在B发生的情况下，A发生的概率

条件概率(Conditional probability)

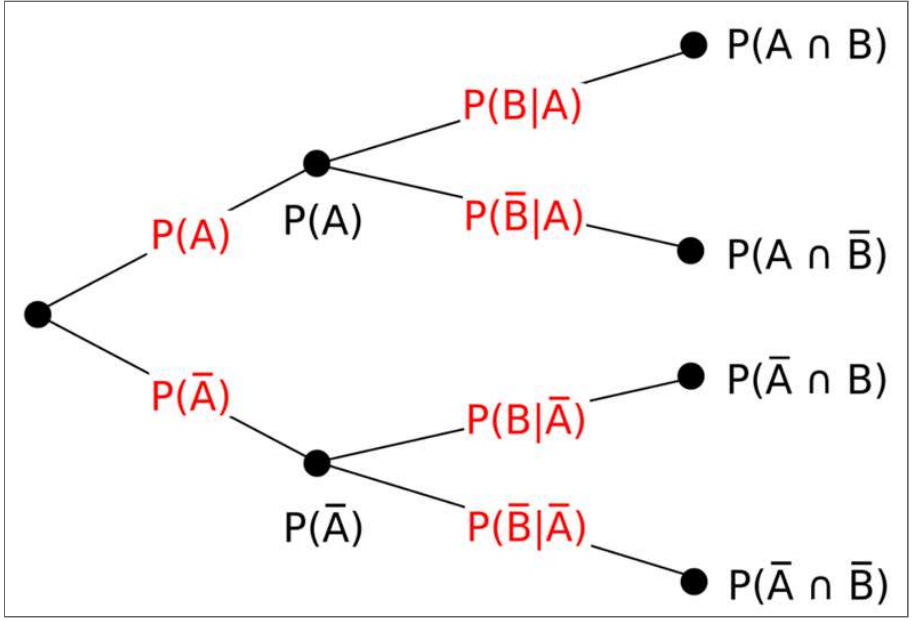
- 若两个事件彼此不独立，在一个事件已经发生的情况下，对另一个事件发生的概率的衡量。

- 如果感兴趣的事件是A，而事件B已知或假定已经发生，“给定B的A的条件概率”，或“B条件下A的概率”，通常写为：

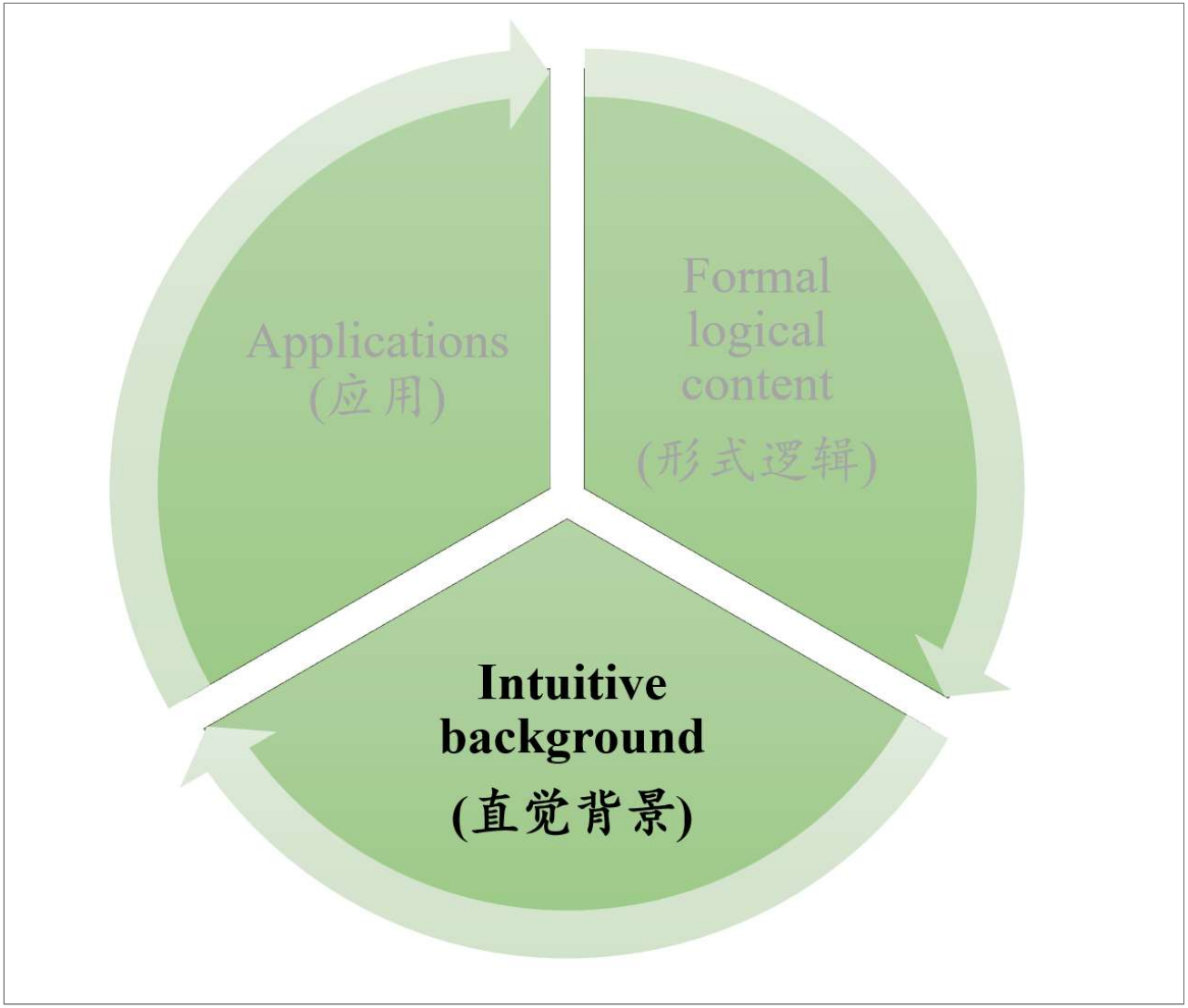
$$p(A|B) = p(A \cap B)/p(B)$$

贝叶斯定理(Bayes' theorem)

- $p(B|A) = \frac{p(A \cap B)}{p(A)}$
- $p(A|B) = \frac{p(A \cap B)}{p(B)}$
- $\rightarrow p(A|B) = \frac{p(B|A)p(A)}{p(B)}$


P(A ∩ B)
P(B|A)
P(A)
P(A)
P(B̄|A)
P(A ∩ B̄)
P(Ā)
P(Ā ∩ B)
P(B|Ā)
P(Ā)
P(B̄|Ā)
P(Ā ∩ B̄)

模型(Model)


Applications
(应用)
Formal
logical
content
(形式逻辑)
Intuitive
background
(直觉背景)

不同的模型可以用来解释相同的现象

The manner in which mathematical theories are applied ***does not depend on*** *preconceived ideas; it is a purposeful technique depending on, and changing with,* ***experience****.*

I have a 100% chance of earning $3000

I have a 75% chance of earning $4000, but a 25% chance of none

你的选择是什么?

I have a 100% chance of **losing** $3000

I have a 75% chance of **losing** $4000, but a 25% chance of none

你的选择是什么?

你可以将此结论拓展到总体吗?

你可以解释这个结果吗?

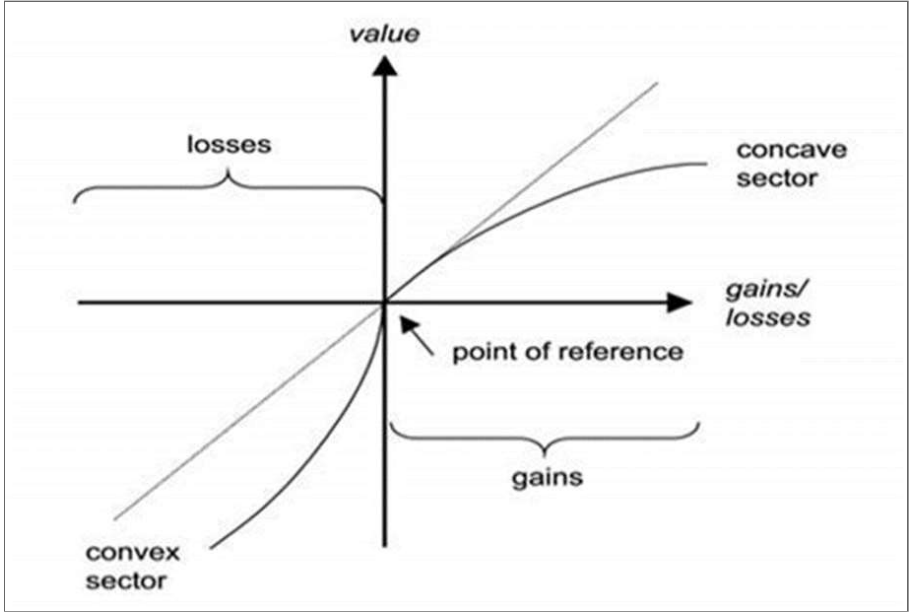
模型可以帮助我们进行解释吗?

模型1：Rational model

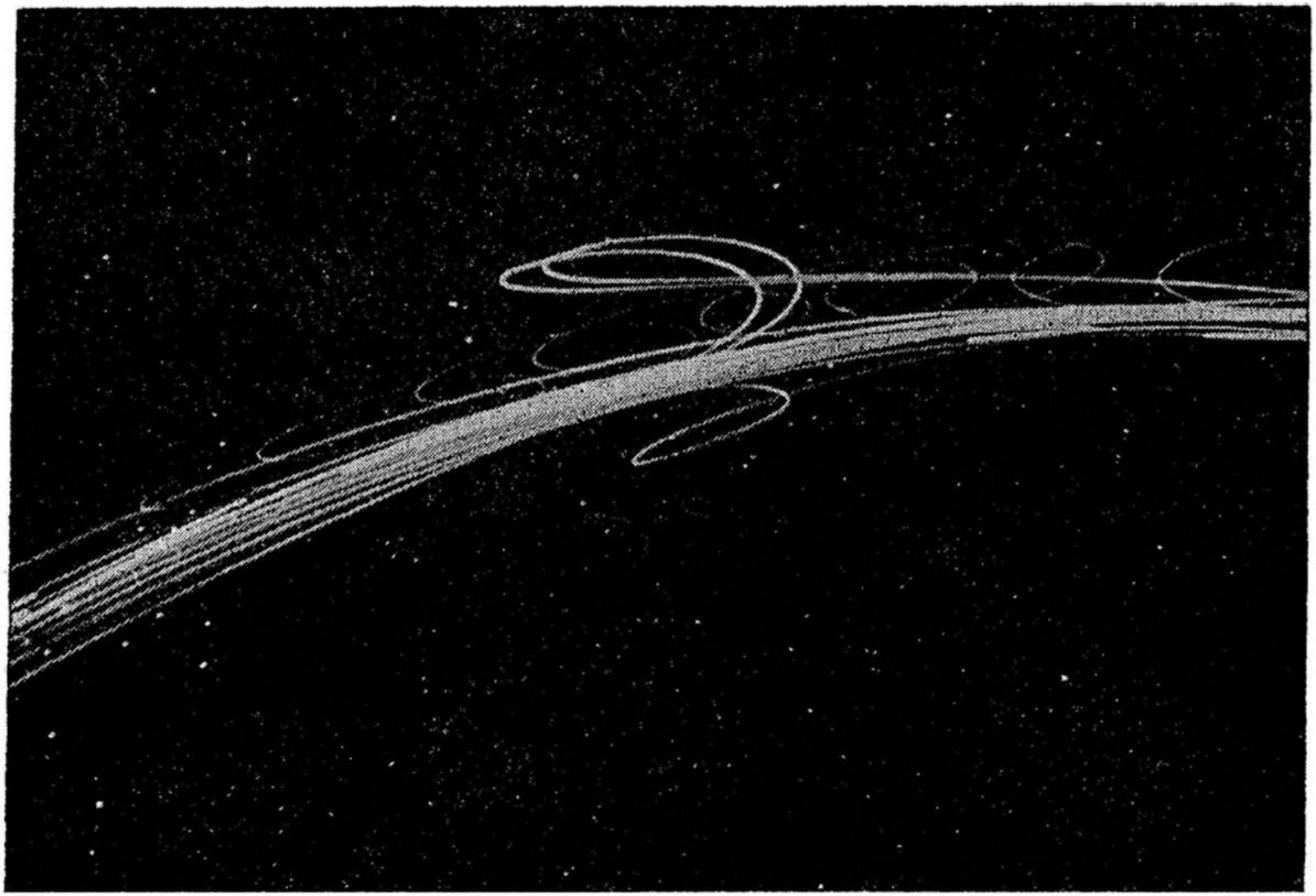
- 理性经济人假设
- 决策者基于每个选项的期望值进行决策

模型2：Prospect theory

- 有限理性假设
- 决策者厌恶损失，偏好风险

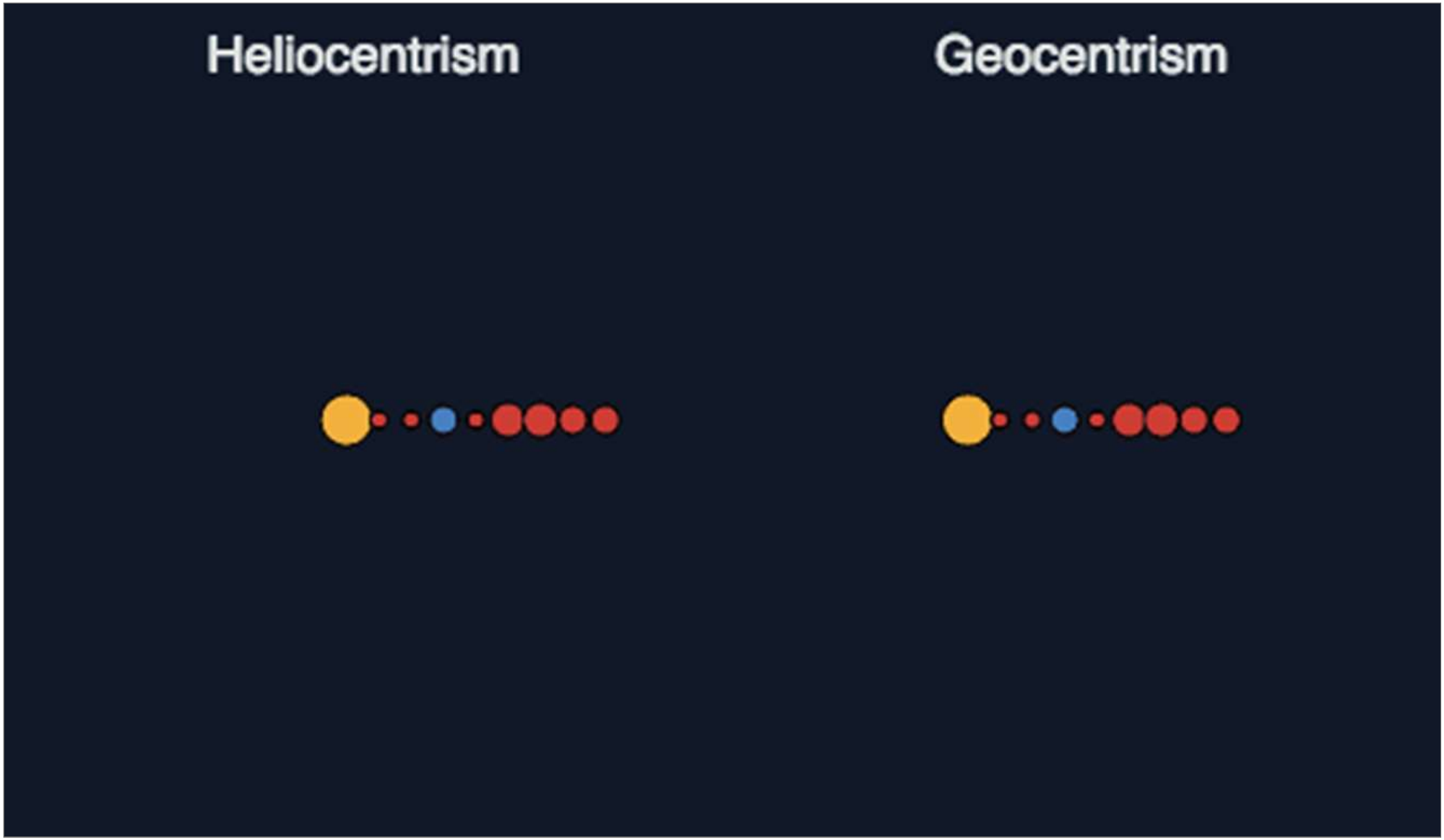

value
losses
concave
sector
gains/
losses
point of reference
gains
convex
sector

你还知道那些模型（不限于心理学）？



这是某一行星移动的轨迹，你能尝试解释它吗

(图片来源： *Proceedings of the Royal Society of London****)***


Heliocentrism
Geocentrism

日心说(Heliocentrism) vs 地心说(Geocentrism)

哪一个模型是正确的呢？日心说(Heliocentrism) 或 地心说(Geocentrism)

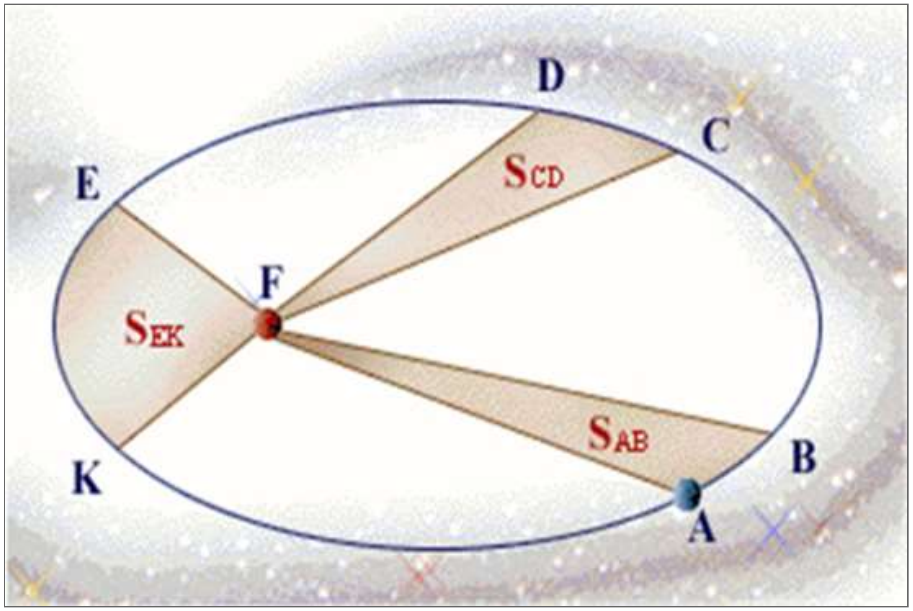
日心说和地心说都假设行星的运动轨迹是圆周，但是实际上行星的运动轨迹是椭圆，所以两个模型都存在问题

开普勒三大定律

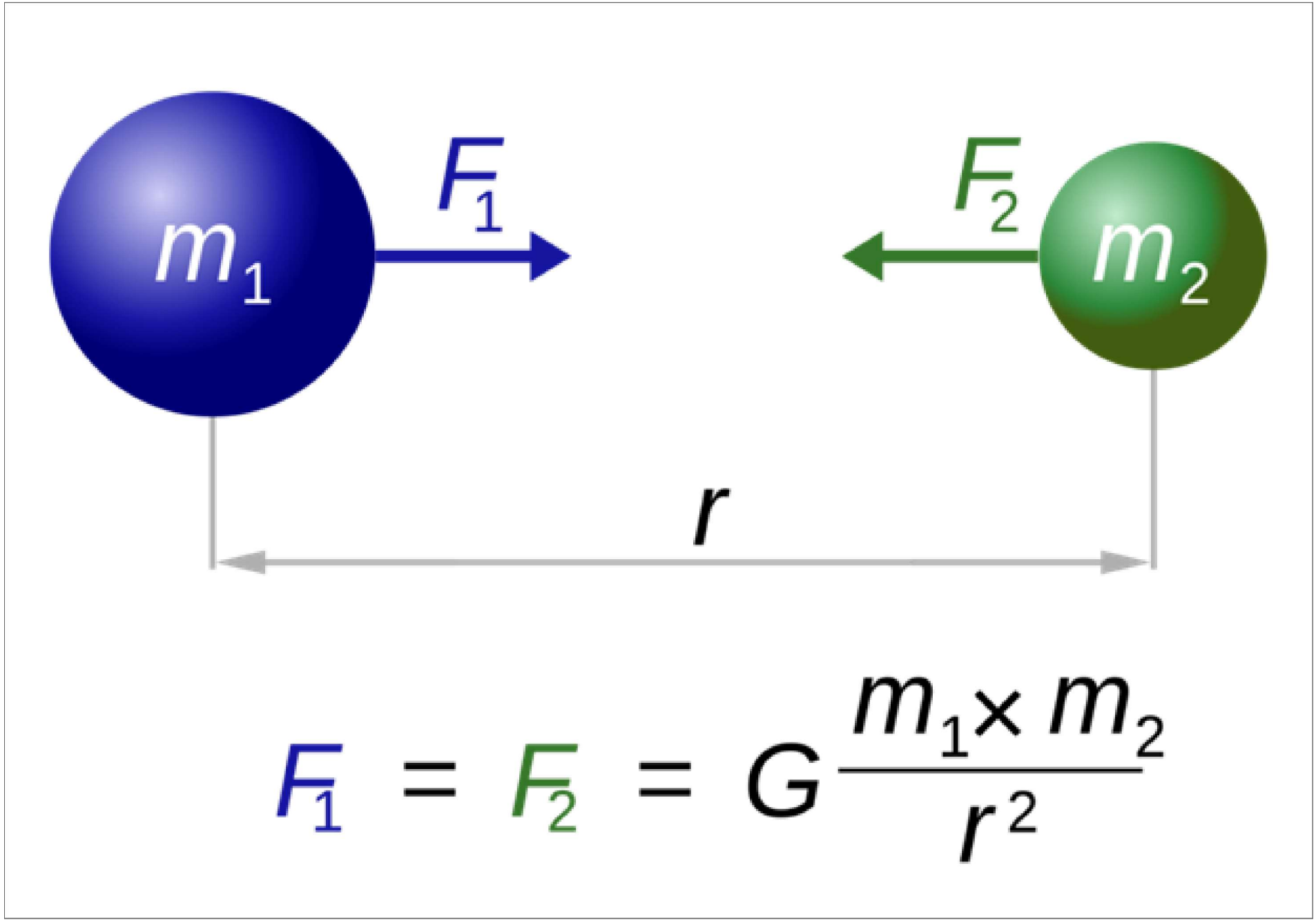
-轨道定律(The Law of Ellipses)

-面积定律(The Law of Equal Areas)

-周期定律(The Law of Harmonies)

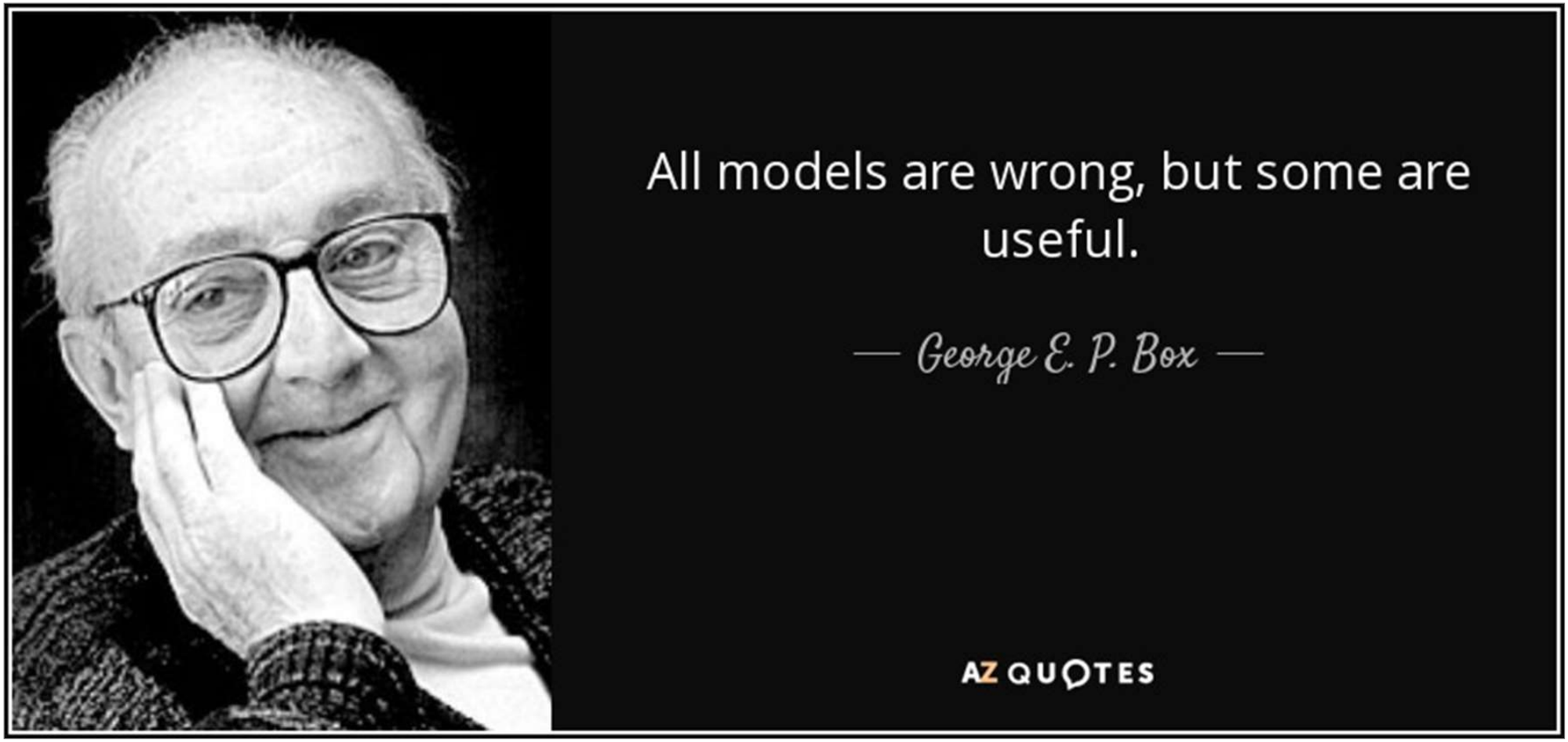


万有引力定理

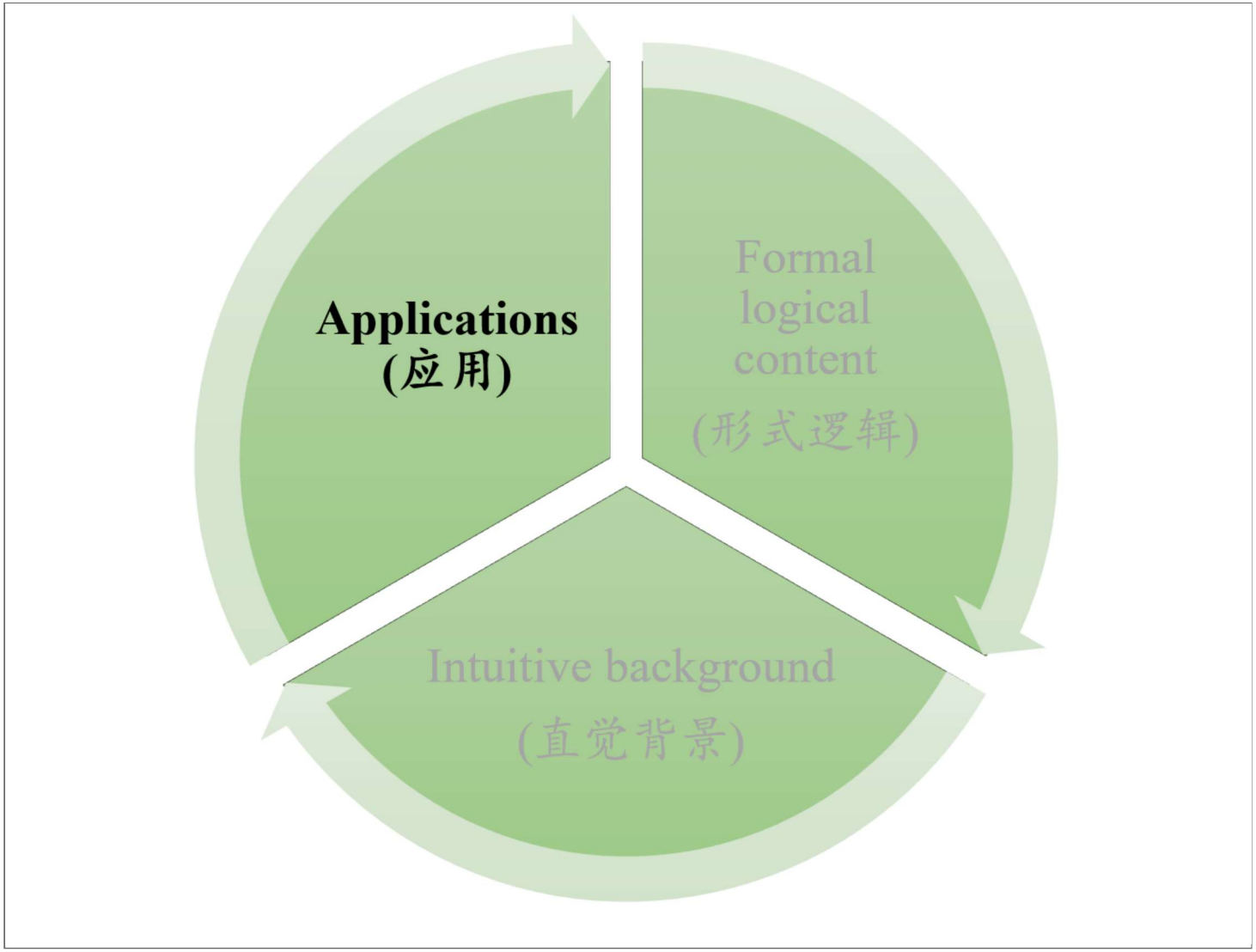


1846年，天文学家发现了海王星，它是围绕我们太阳运行的第八颗行星。
这一发现是基于对其预测位置的数学计算，而这一预测是由天王星的轨道上观察到的扰动造成的，使用的是万有引力定律。

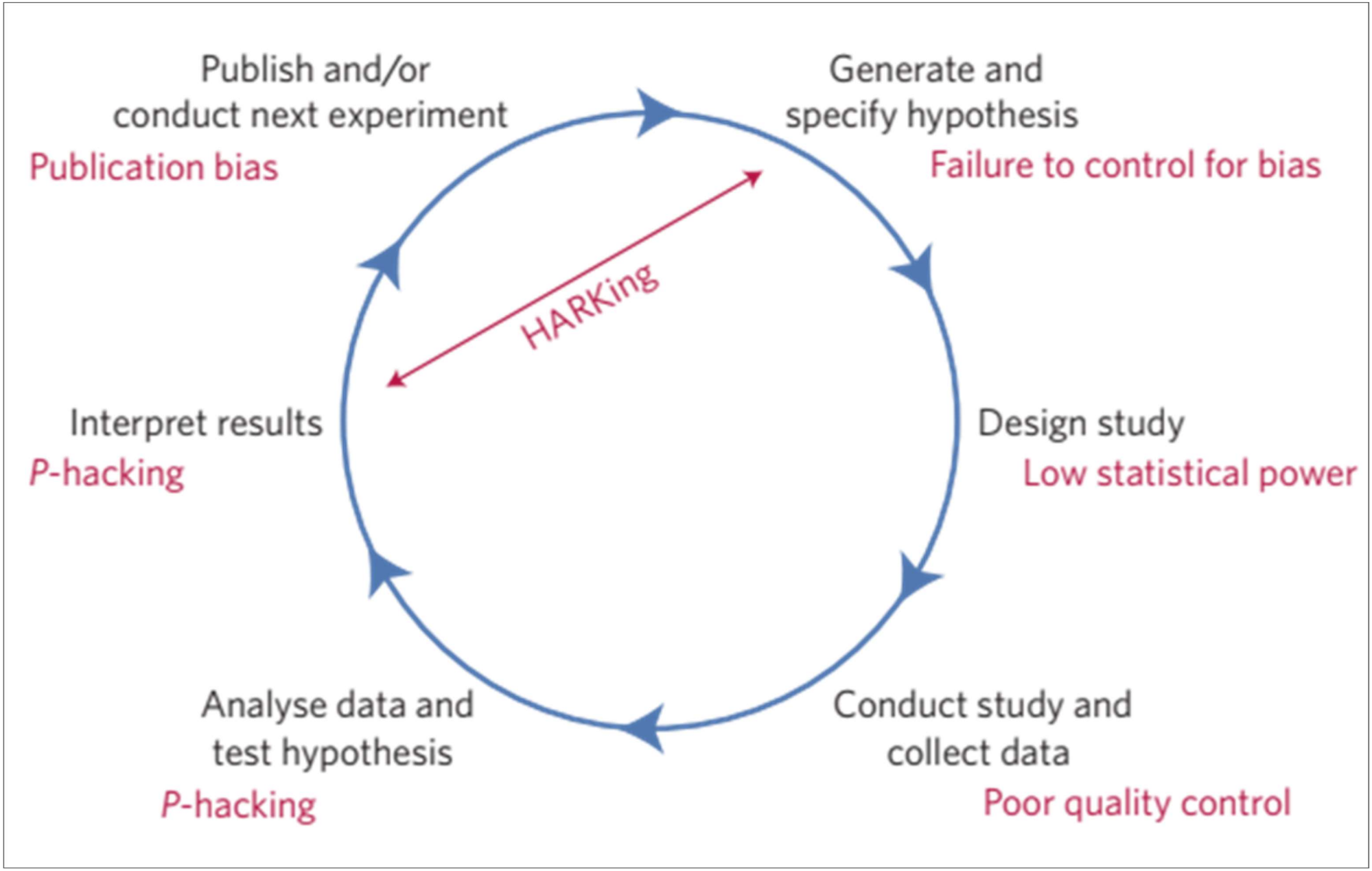
万有引力定律是正确的模型吗?


All models are wrong, but some are useful.
— George E. P. Box —
AZ QUOTES

研究中的统计(Statistics)



研究的基本流程


Publish and/or
conduct next experiment
Publication bias
Generate and
specify hypothesis
Failure to control for bias
HARKing
Interpret results
P-hacking
Design study
Low statistical power
Analyse data and
test hypothesis
P-hacking
Conduct study and
collect data
Poor quality control

(Munafò et al., 2017. Nat Hum Behav)

研究设计

描述性研究(Descriptive Research)

相关性研究(Relational or Correlational Research)

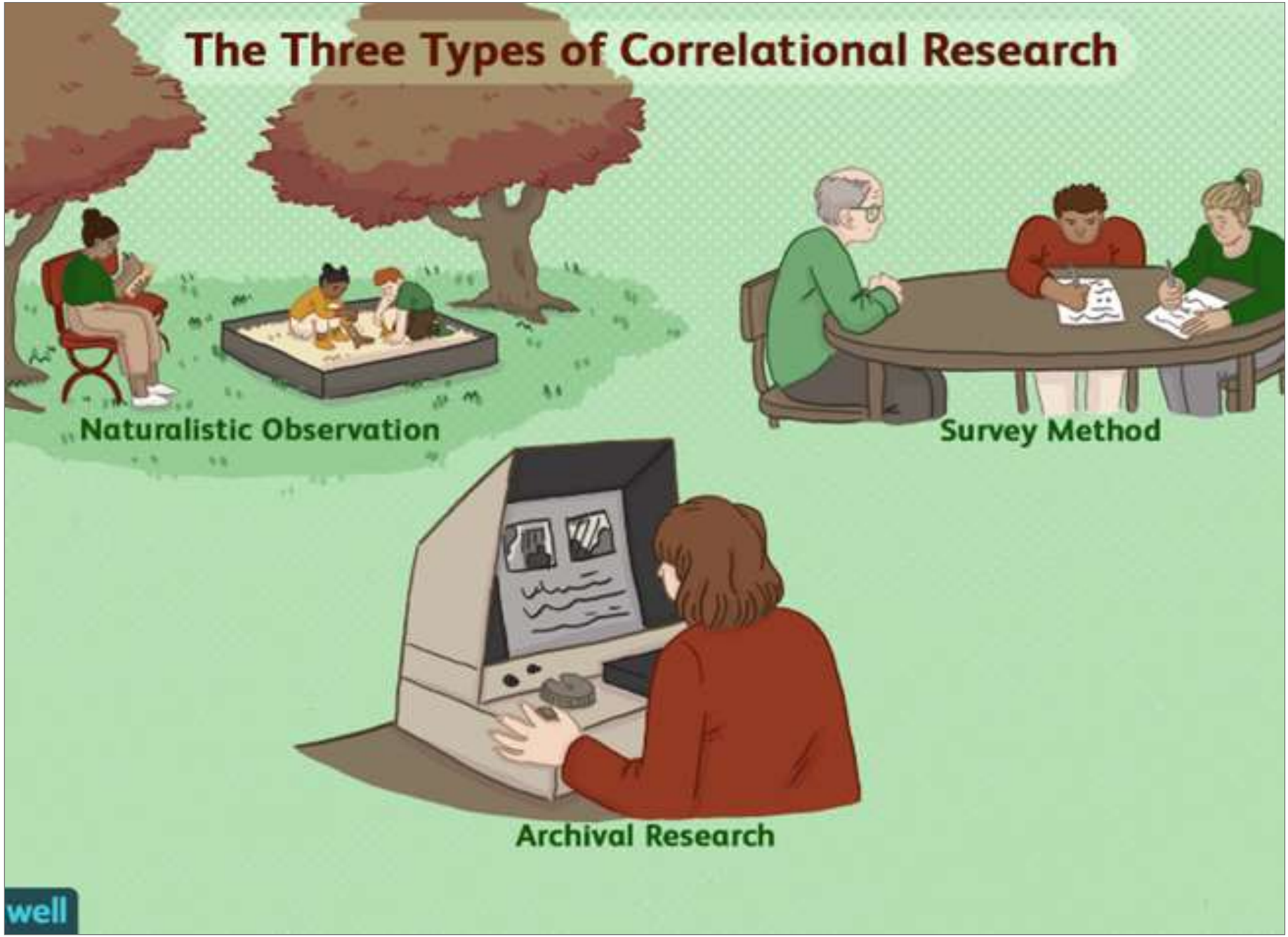
实验研究(Causal or Experimental Research)

描述性研究

- 描绘已经存在的、在一个群体或人群中的情况;

- 不是为了测量效果，而只是为了描述效果。

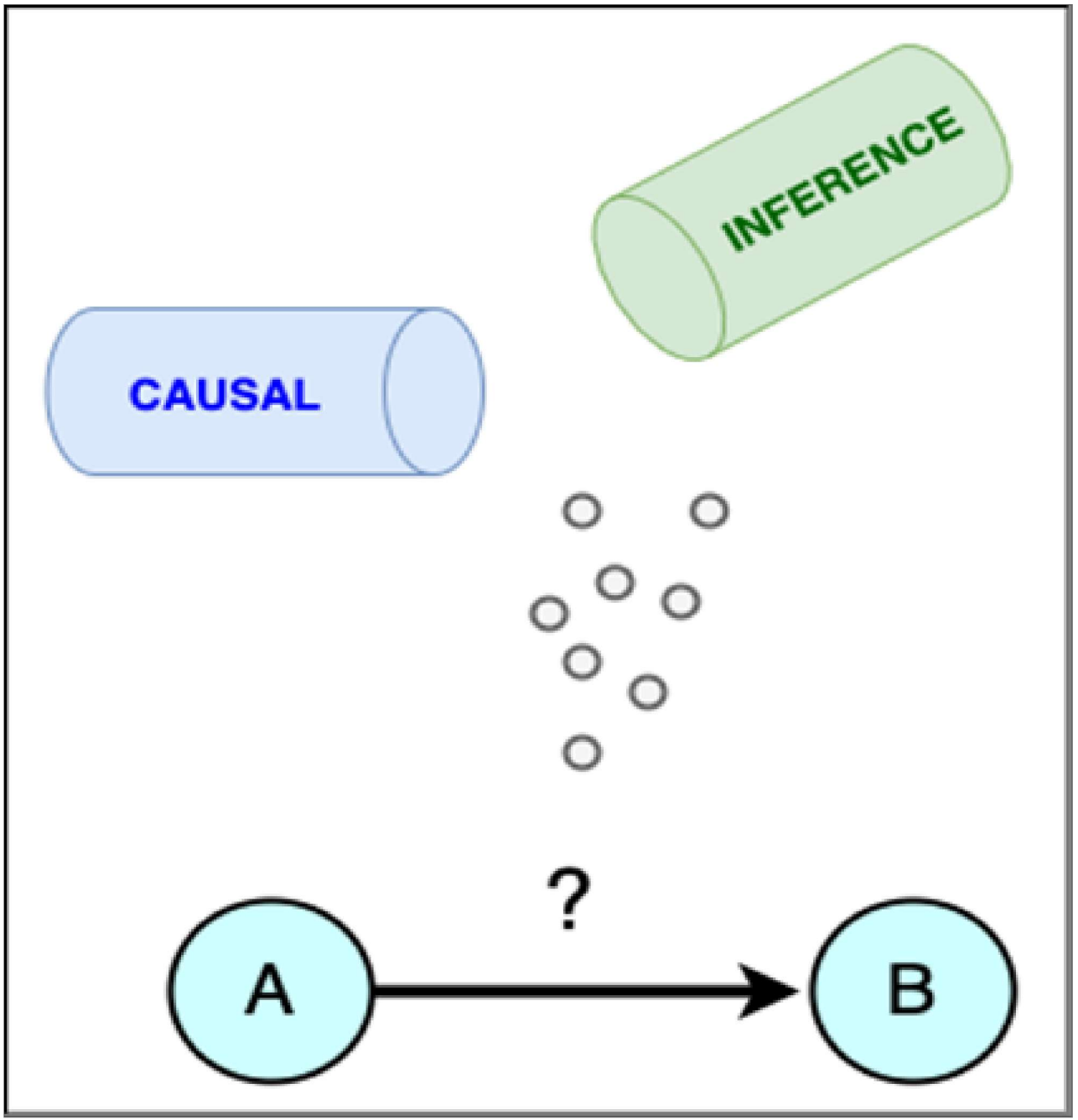
相关性研究


The Three Types of Correlational Research
Naturalistic Observation
Survey Method
Archival Research
well

- 调查两个或多个变量之间的联系;

- 被比较的变量一般已经存在于群体或人群中。

实验研究



- 调查一个或多个变量对一个或多个结果变量的影响;

- 确定一个变量是否导致另一个变量发生或变化。

问题

哪一种研究更常见于心理学研究？为什么？

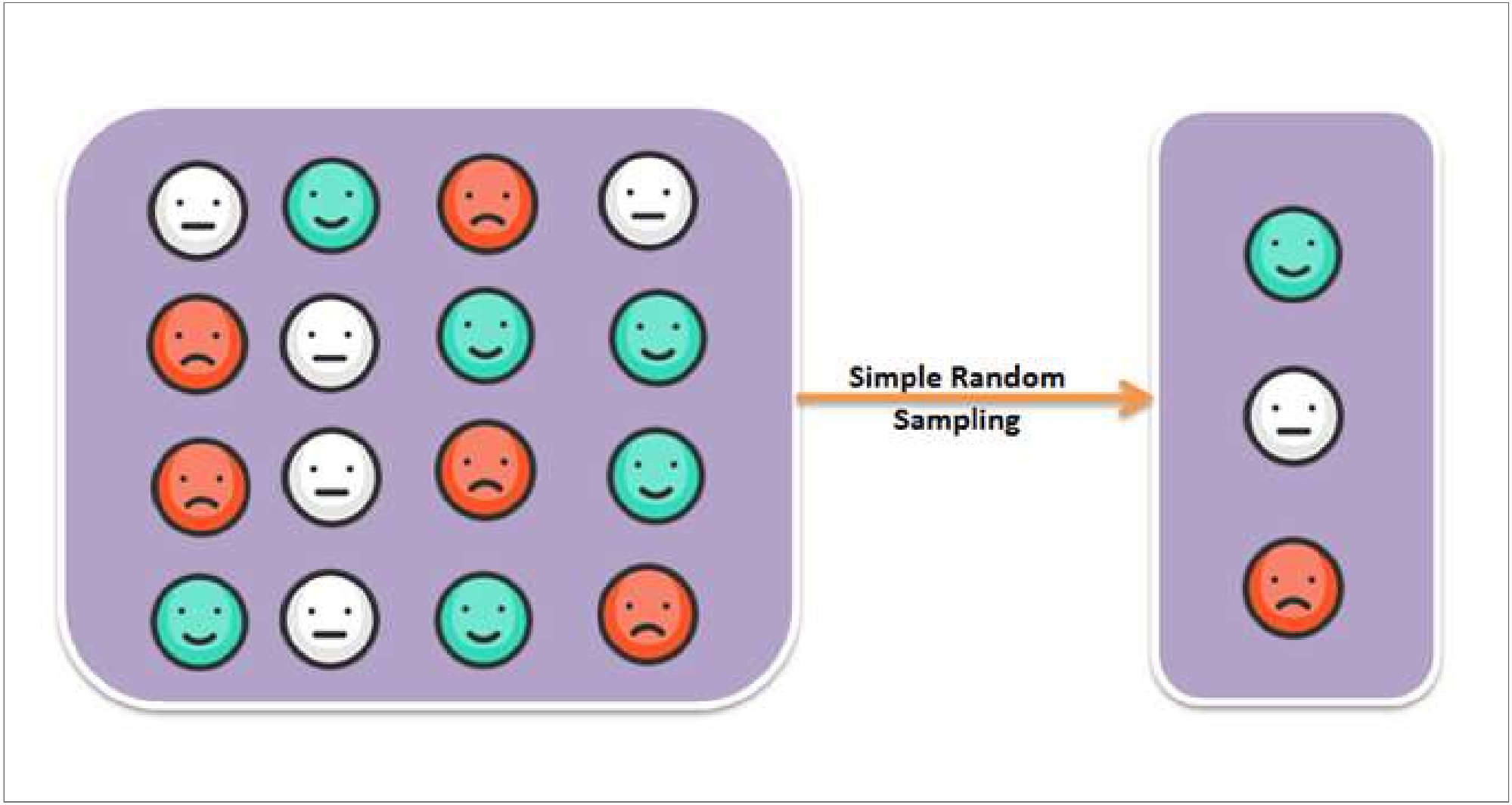
你可以预测你能记住多少东西吗?

搜集资料 (Collection)

操作性定义(Operational definition)

- 理论定义与操作定义。
- 模拟或代表一个概念或理论定义，也被称为构造(construct)。
- 描述定义该概念的操作（程序、行动或过程），要有足够的具体性，以便其他调查者可以复制他们的研究。
- 用具体的、可公开的准备或验证测试的过程来定义系统状态。

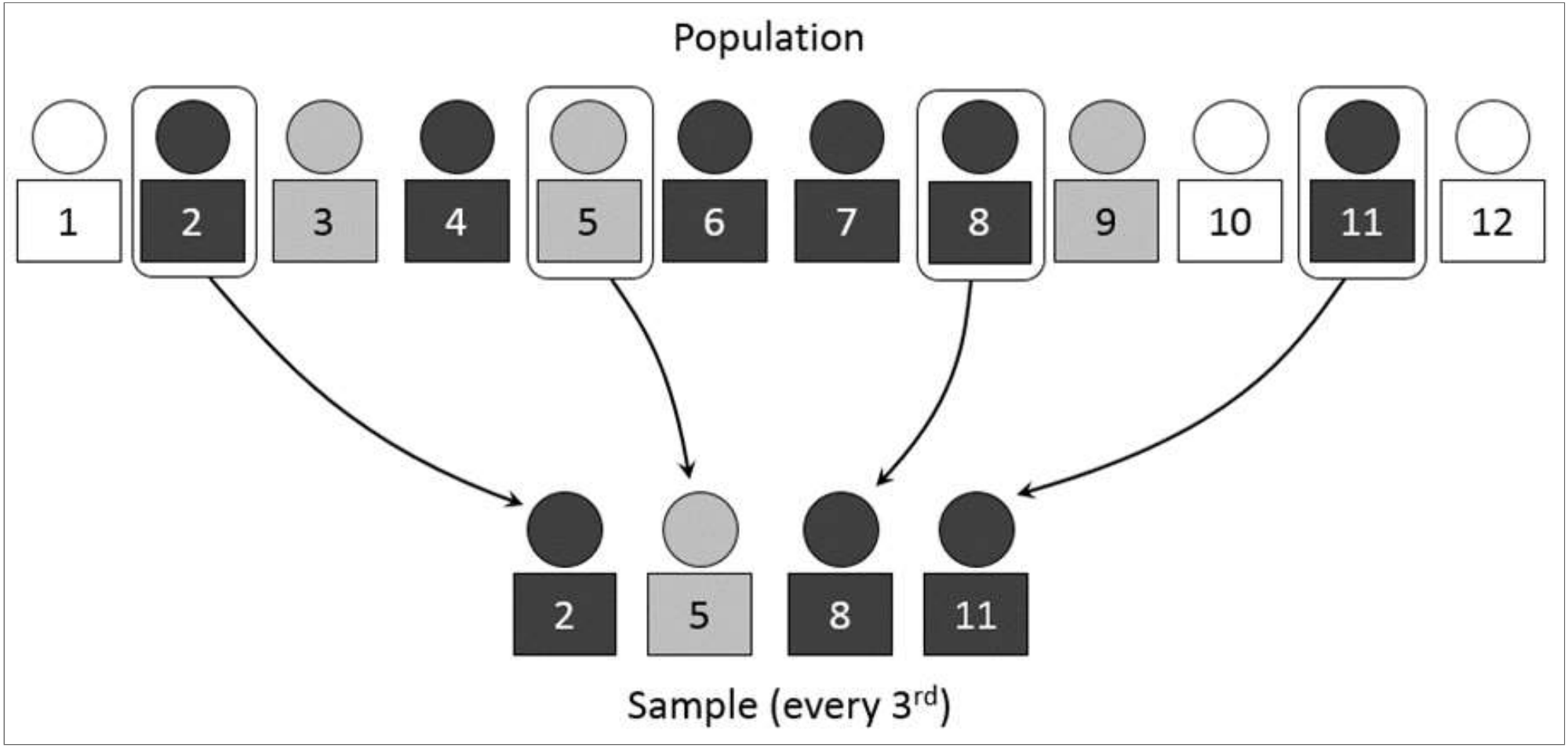
简单随机抽样(Simple random sampling)



- 每个个体的选择完全是偶然的，总体中的每个个体被选中的机会或概率是相等;

- 获得随机样本的一种方法是给总体中的每个个体一个编号，然后用一个随机数字表来决定包括哪些人。

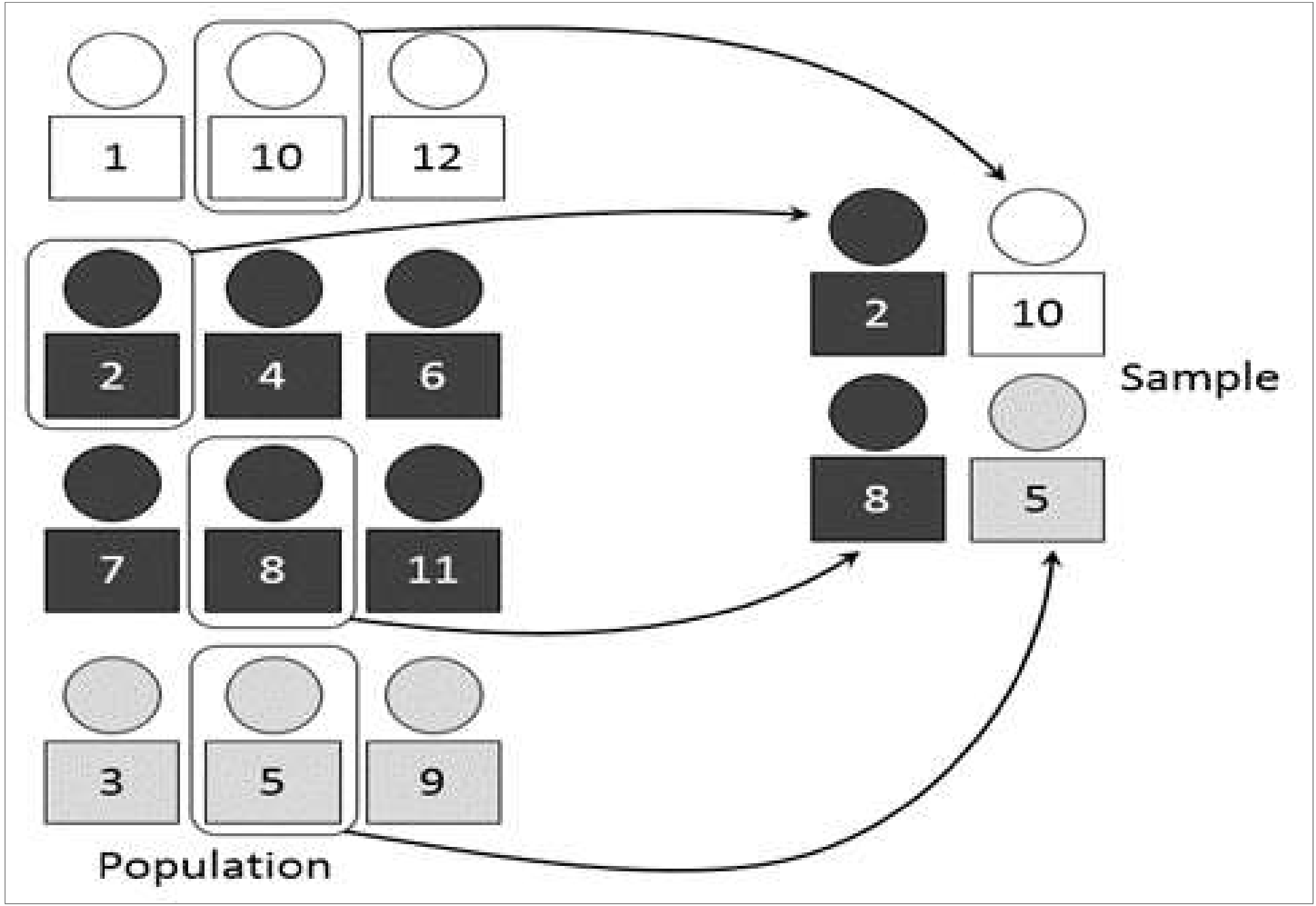
系统抽样(Systematic sampling)



- 从抽样框架中按照规律间隔选择个体;
- 间隔的选择是为了确保足够的样本量;

- 如果你需要从一个大小为x的总体中抽取n个样本，你应该选择每x/n个个体作为样本。

分层抽样(Stratified sampling)

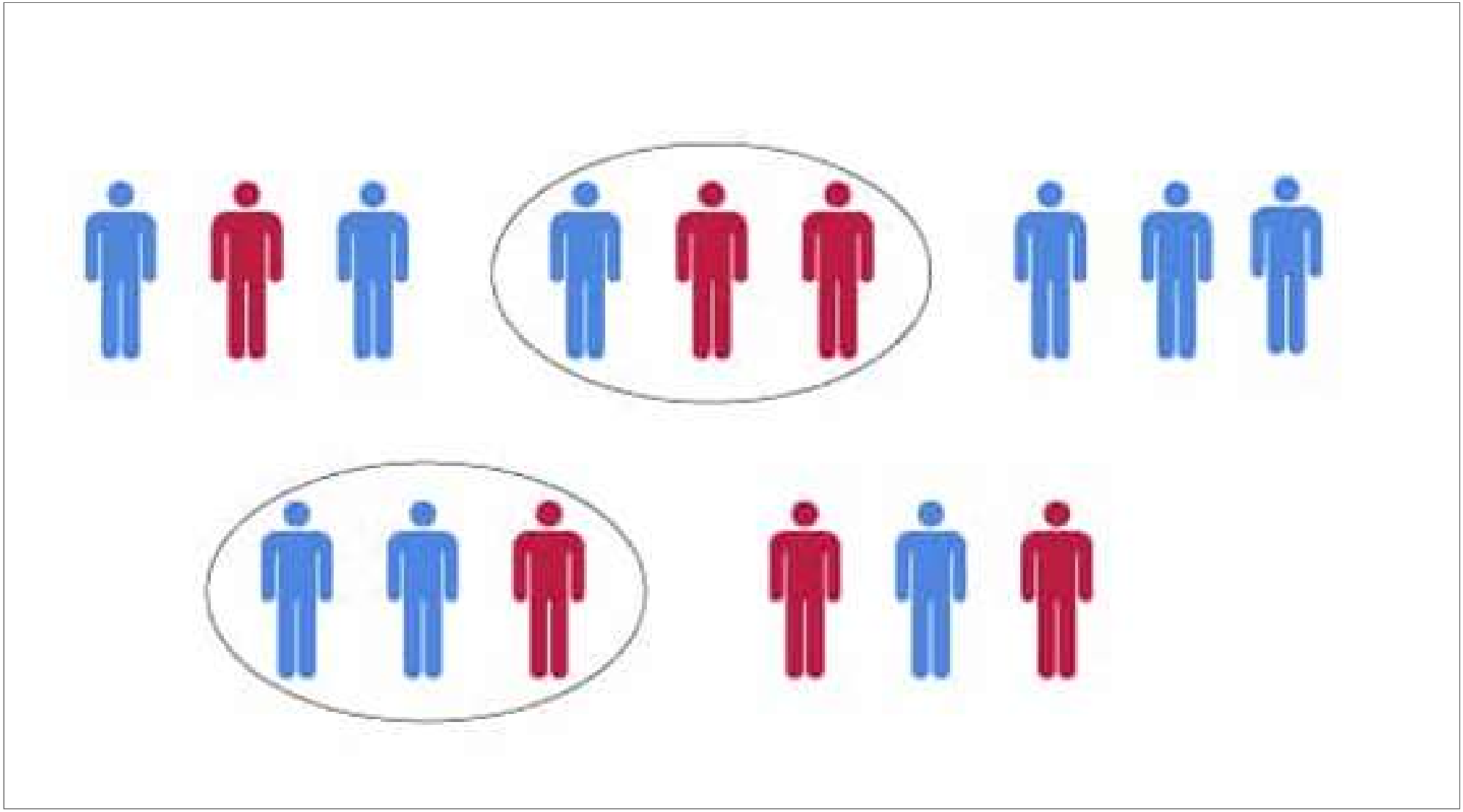


- 当我们测量值在不同的子群体之间有所不同，并且我们想确保所有子群体的代表性时，就会使用这种方法;

- 总体被划分为具有类似特征的层级，不同的层级中的样本存在差异。

- 为了确保样本的代表性，每个层级中抽取合适的样本量来获得研究样本。

分组抽样(Clustered sampling)



- 以总体的子组作为抽样单位，而不是单个个体。

- 人群被分为亚组，称为集群，这些亚组是随机选择纳入研究的。
- 在单阶段整群抽样方法中，将所选整群的所有成员纳入研究。
- 在两阶段整群抽样中，从每一群中随机选择一个个体纳入研究。

非概率抽样方法(Non-Probability Sampling Methods)

- 配额抽样(Quota sampling)
- 判断性或目的性抽样(Judgement /or Purposive Sampling)
- 雪球抽样(Snowball sampling)

什么是记忆？记忆该怎样测量？

艾宾浩斯所使用的抽样方法以及样本是什么?

抽样偏倚(Bias in sampling)

- 在选择样本时,应该考虑5个重要的潜在偏差来源,而不考虑使用的方法。
  - 偏离了预先商定的抽样规则[1]
  - 遗漏了难以接触到的人群中的个体[2]
  - 将被选中的人替换成其他人，比如在难以联系上他们的时候[3]
  - 回复率很低[4]
  - 过时的名单被用作样本框架（例如，如果它不包括最近搬到某个地区的人）。[5]

变量(Variable)

- 常量与变量
  - Constant and variable
- 自变量和因变量
  - Independent variable and dependent variable
- 连续变量和离散变量
  - Continuous variable and discrete variable ##

- 常量 (Constant)
  - 在整个实验过程中保持不变的变量，而其他变量可能发生变化。

- 变量(Variable)
  - 任何可以采取不同值的特征，如身高、年龄、温度或考试分数。

Constant and variable vary with research question and design
同一个变量在不同研究中角色可能会变化

变量(Variable)

- 自变量(Independent variable)
  - 自变量是导致变化的原因
  - 它的值与研究中的其他变量无关
- 因变量(Dependent variable)
  - 因变量是变化的结果
  - 其数值取决于自变量的变化

变量(Variable)

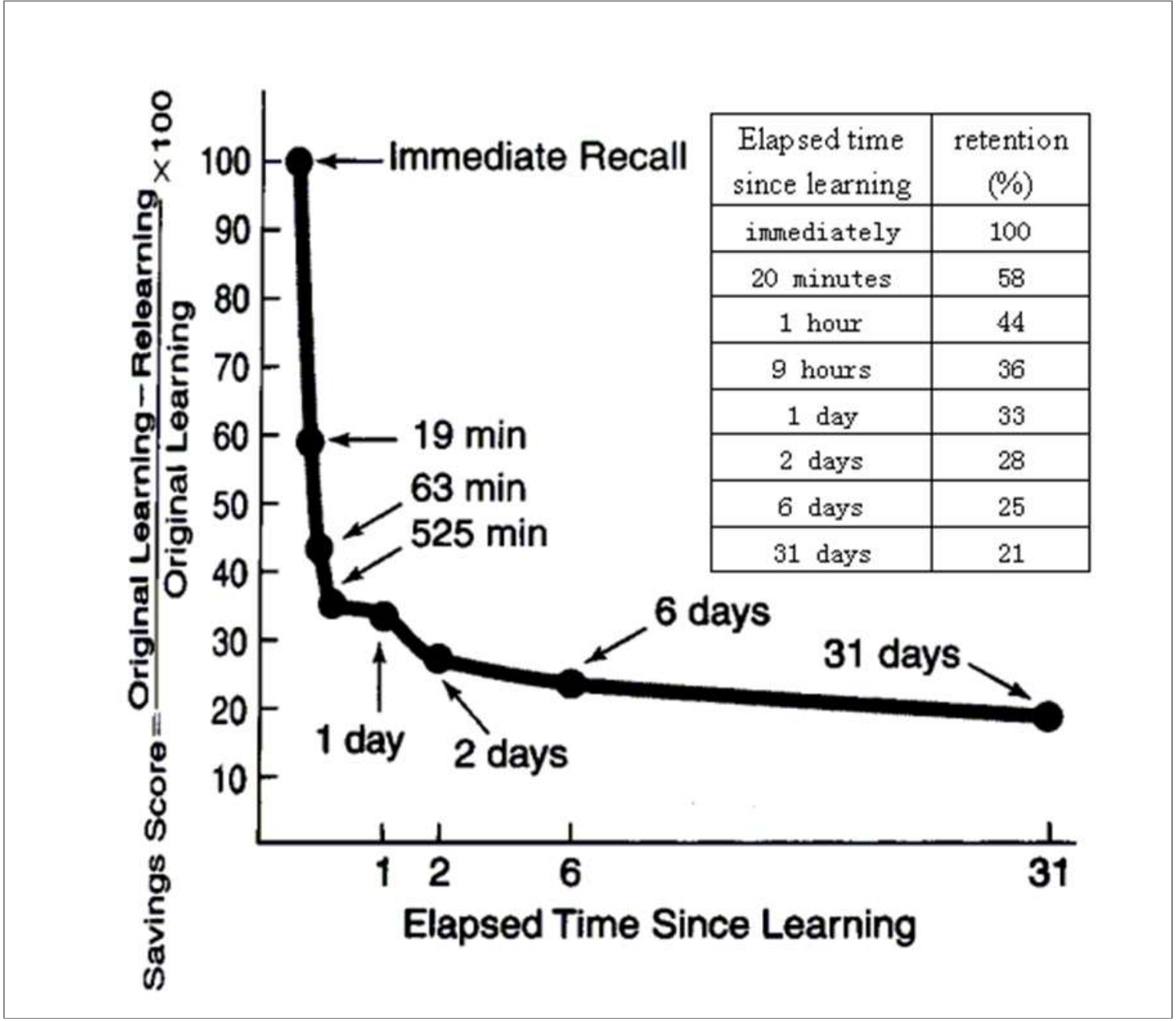
- 连续变量(Continuous variable)
  - 一个变量可以有无限多的可能值

- 离散变量(Discrete variable)
  - 一个变量只能取有限数量的值
  - 所有的定性变量都是离散的，有些定量变量是离散的
  - 有时，一个具有足够离散值的变量在实际应用中可能被认为是连续的

记住的音节个数是什么变量?

预处理 (Preprocessing)

数据可视化(Data visualization)



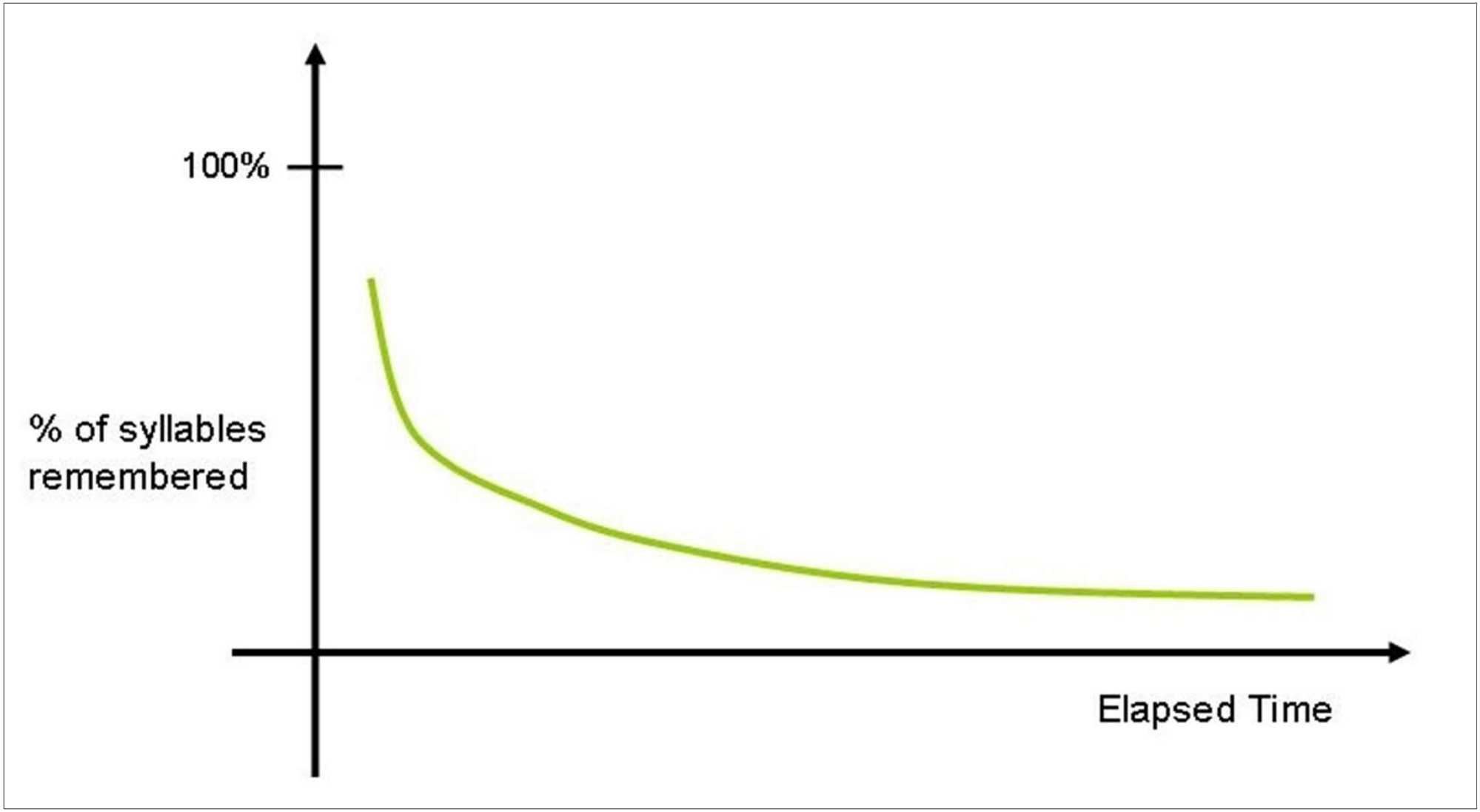
Elapsed time since learning	retention (%)
immediately	100
20 minutes	58
1 hour	44
9 hours	36
1 day	33
2 days	28
6 days	25
31 days	21

The Ebbinghaus Forgetting Curve

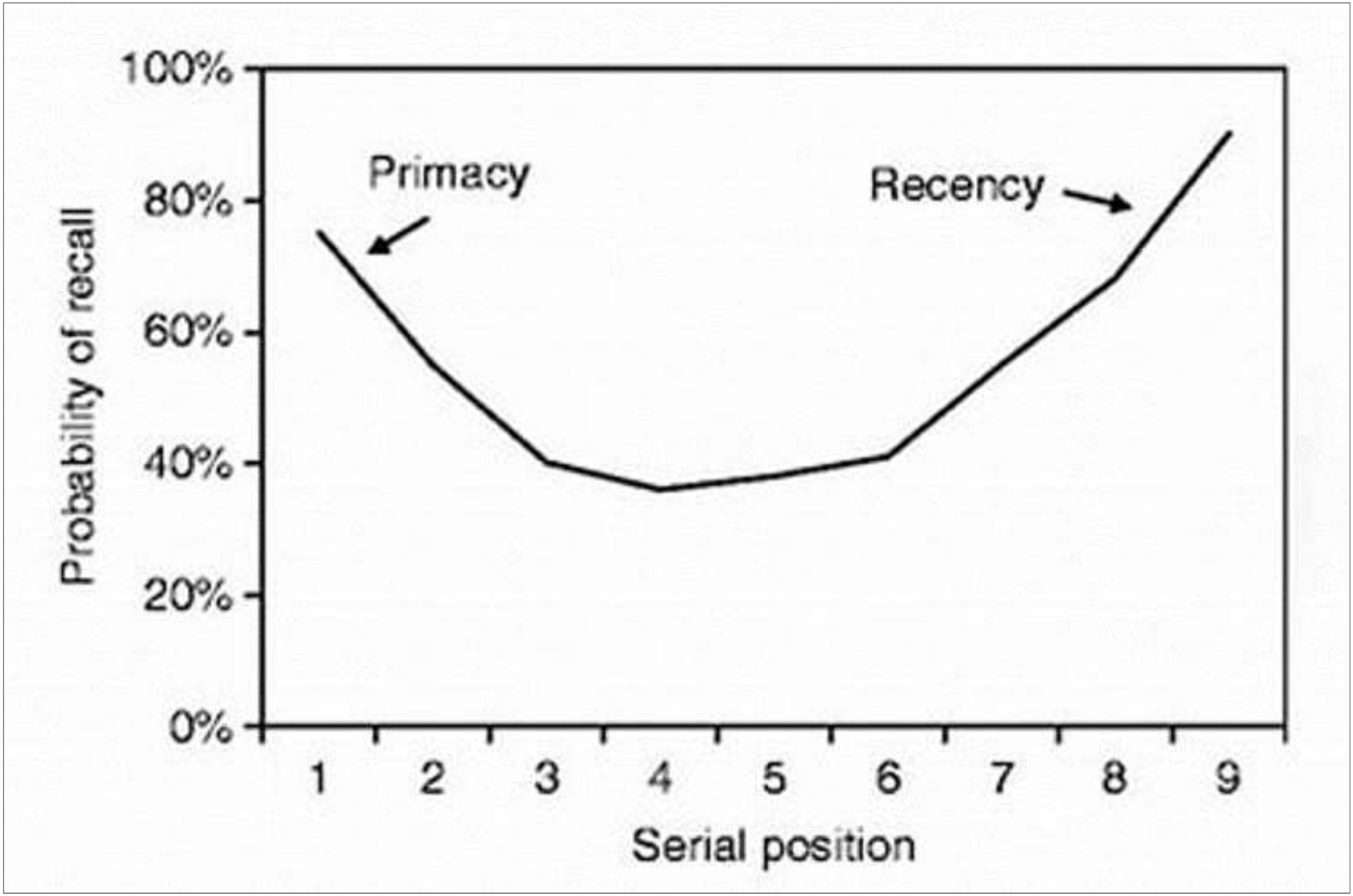
统计分析与推断 (Analyze)

- 模型(Model)
  - 采用何种模型，为什么?
- 结果(Results–a bunch of numbers)
  - 结果是有效的吗?
- 推断(Inference–binary or quantitative estimate)
  - 可以从数字中推断出什么?
- 做出决策(Decision-making)

以下哪种模型是正确的?

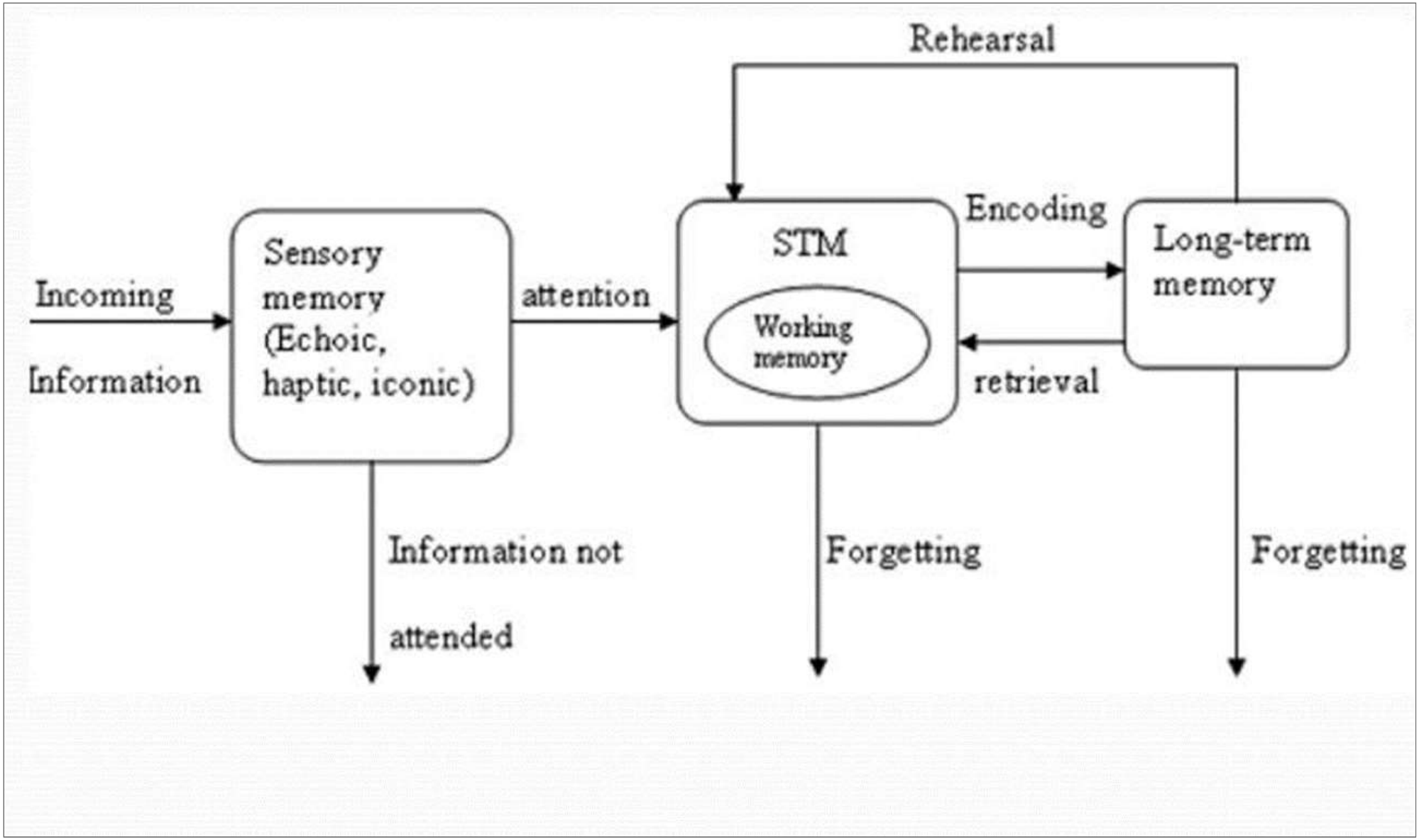


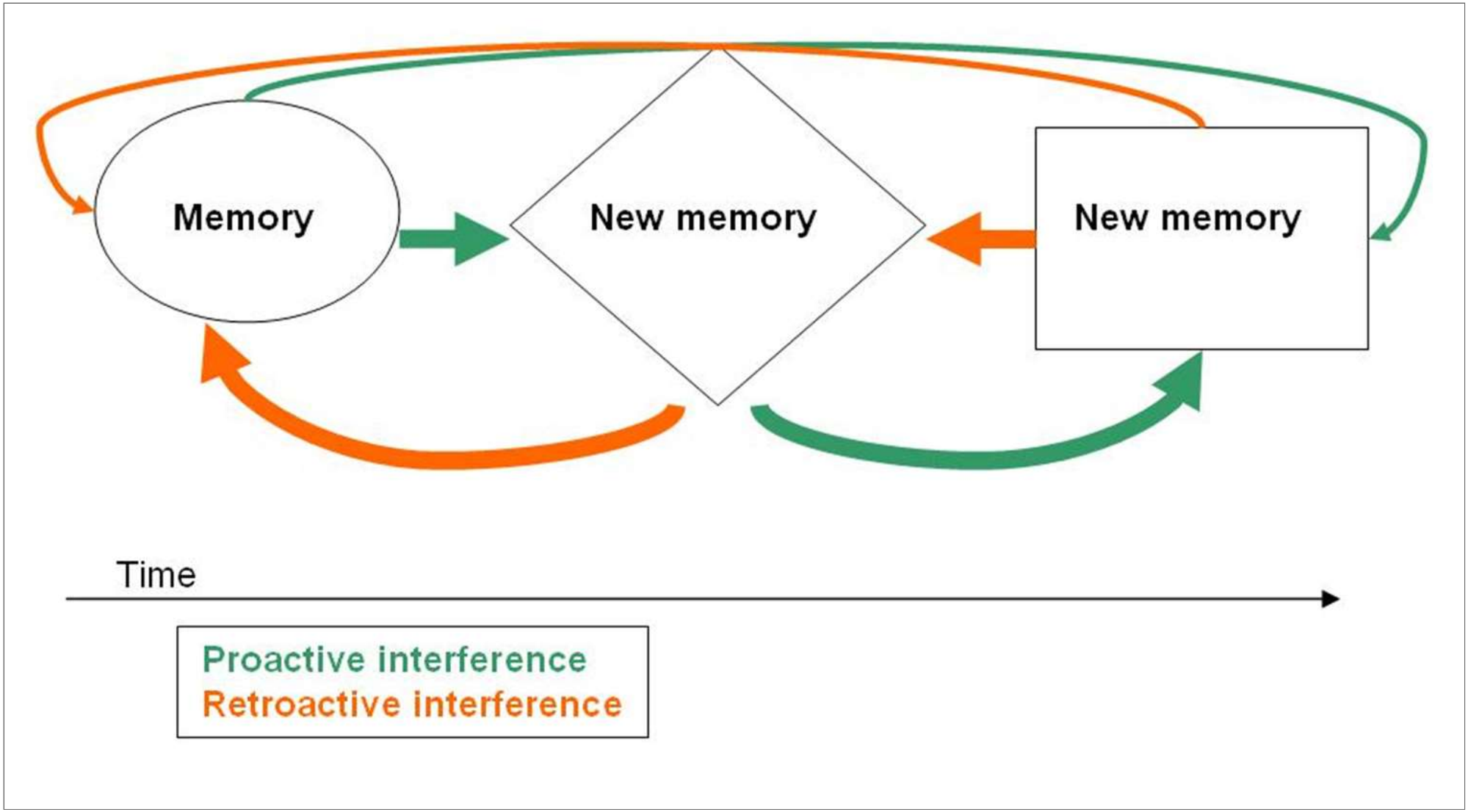
艾宾浩斯遗忘曲线(The Ebbinghaus Forgetting Curve)


100%
80%
60%
40%
20%
0%
Probability of recall
Primacy
Recency
1
2
3
4
5
6
7
8
9
Serial position

系列位置曲线(The serial position curve)

如何解释结果?


Rehearsal
Incoming
Information
Sensory
memory
(Echoic,
haptic, iconic)
attention
STM
Working
memory
Encoding
retrieval
Long-term
memory
Information not
attended
Forgetting
Forgetting


Memory
New memory
New memory
Time
Proactive interference
Retroactive interference

总结(Summary)

- **概率**

1. Any pre-agreed sampling rules are deviated from
2. People in hard-to-reach groups are omitted
3. Selected individuals are replaced with others, for example if they are difficult to contact
4. There are low response rates
5. An out-of-date list is used as the sample frame (for example, if it excludes people who have recently moved to an area)

- **模型**
- **统计**

Error ×