Python与机器学习

——机器学习简介

华算科技 黄老师
2022年2月22日

华算科技
HUASUAN TECHNOLOGY

您的单位现在如何将机器学习应用于化学当中？（单选）

A. 预测分子的性质/活性
B. 反应预测/逆合成分析
C. 解释实验结果
D. 其它的化学应用
E. 不使用机器学习

您的单位现在如何将机器学习应用于化学当中？（单选）

A. 预测分子的性质/活性	36%
B. 反应预测/逆合成分析	7%
C. 解释实验结果	16%
D. 其它的化学应用	5%
E. 不使用机器学习	36%

数据来源：DASSAULT SYSTÈMES

目录

1. 机器学习是什么

2. 机器学习与化学研究

3. 机器学习库

目录

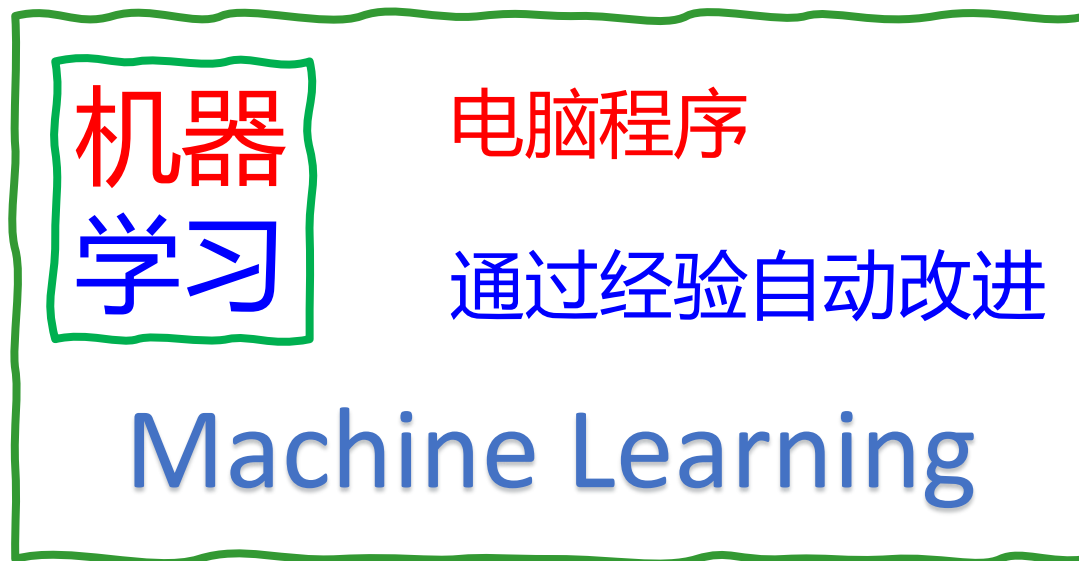1. 机器学习是什么

2. 机器学习与化学研究

3. 机器学习库

石头剪刀布

规则

数据

传统编程方式

结果

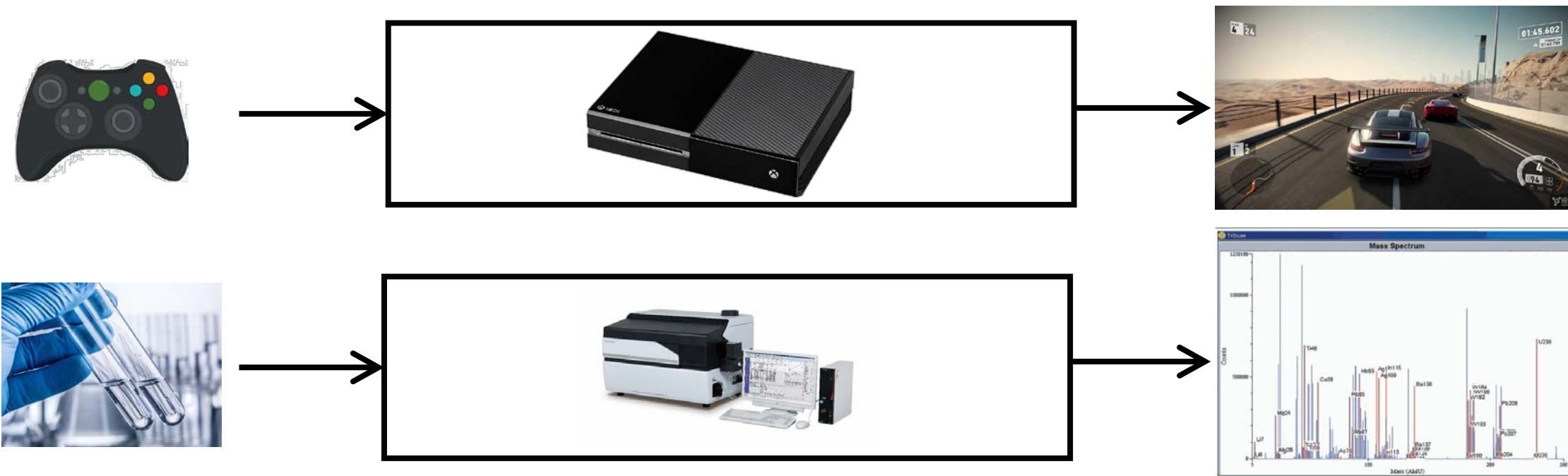Concerned with the question of how to construct computer programs that automatically improve with experience.
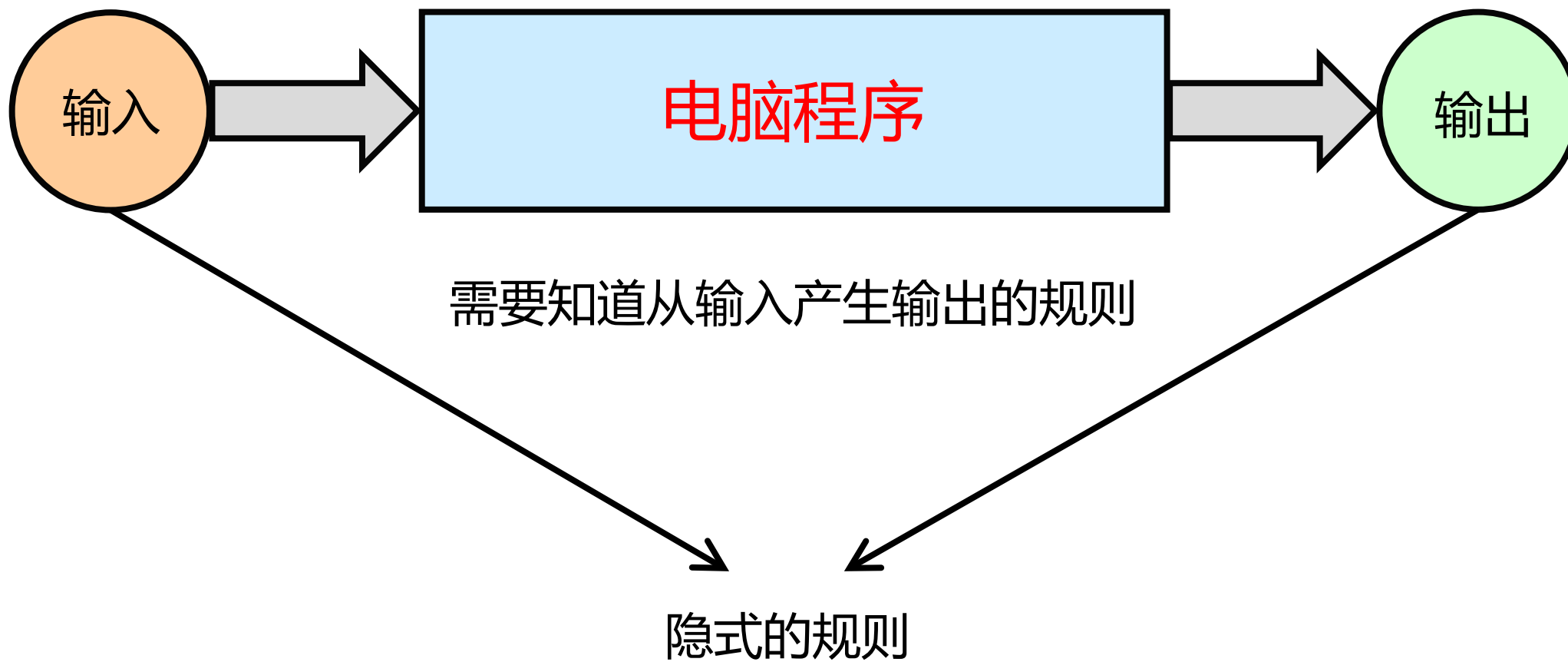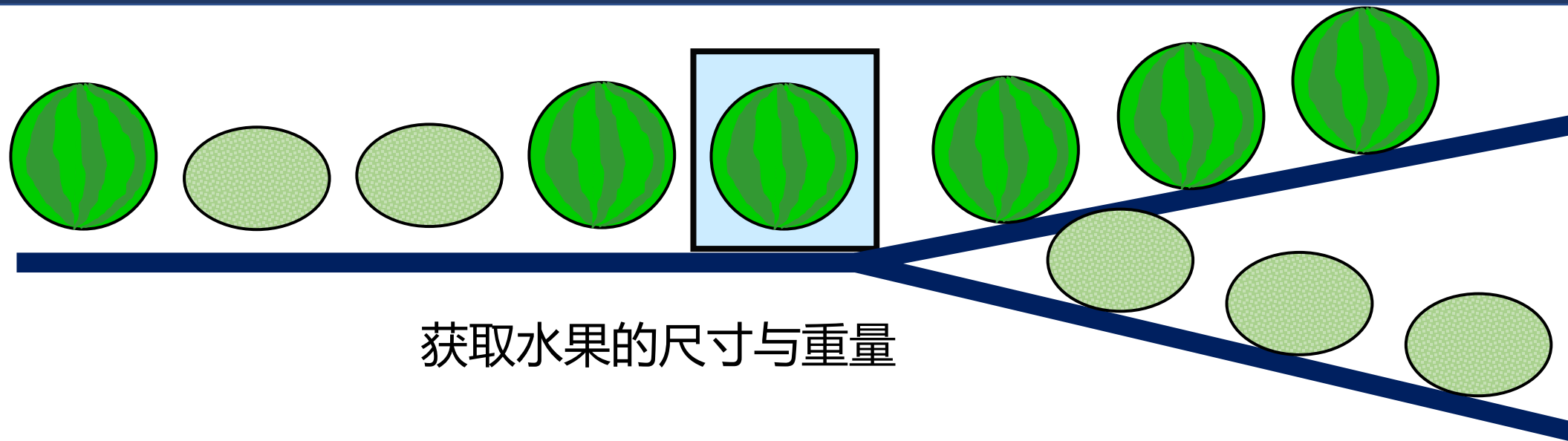
——Tom Mitchell
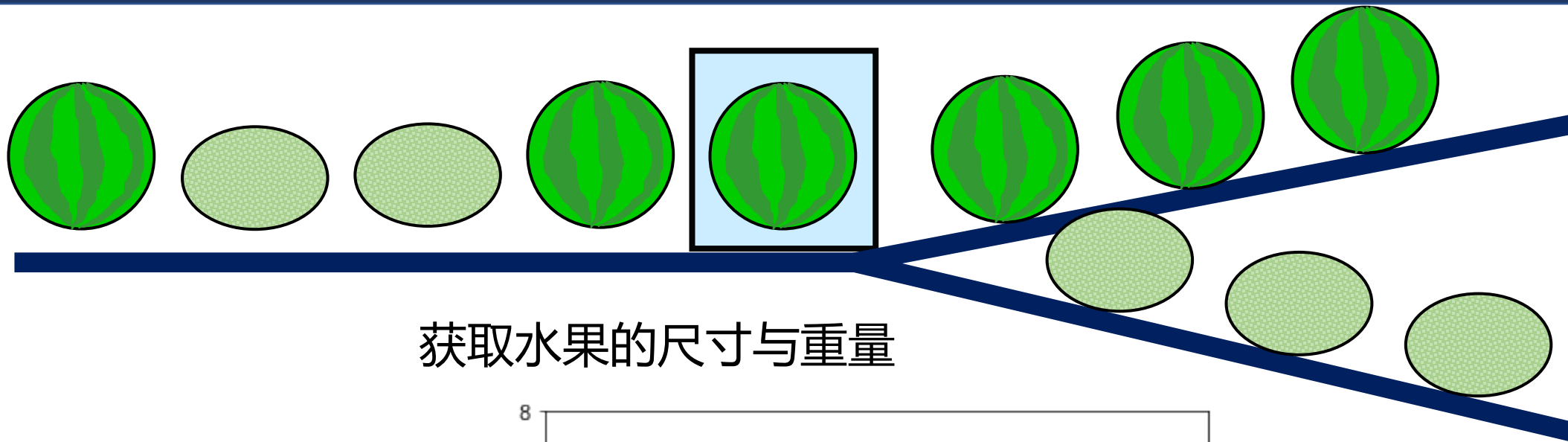
机器学习   电脑程序

通过经验自动改进

Machine Learning

电脑程序？

电脑程序即为处理输入获得输出的方式

输入 → 电脑程序 → 输出
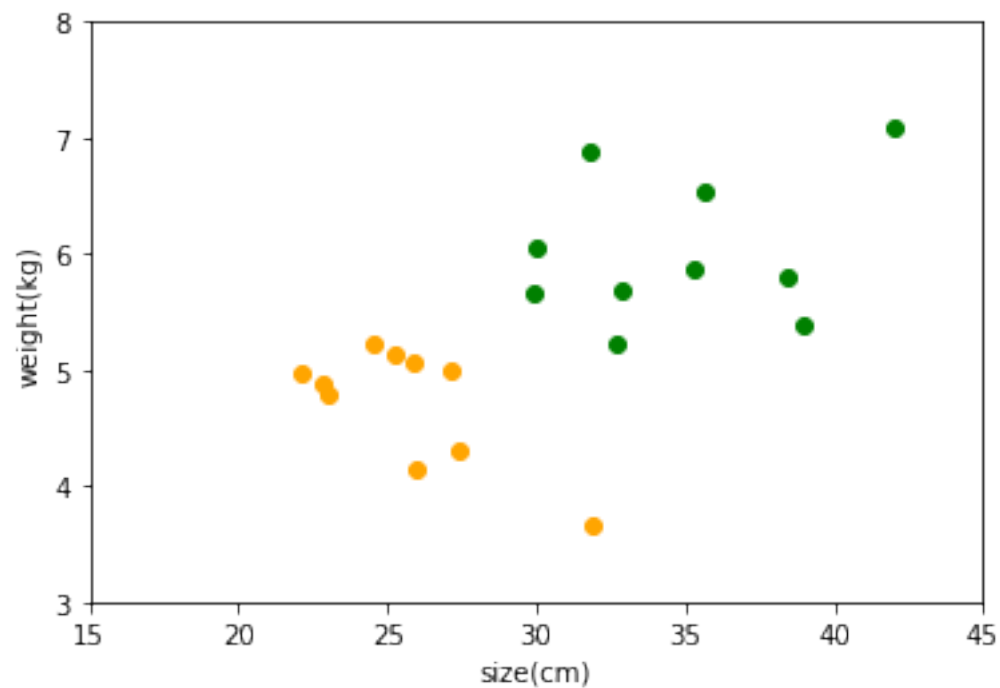
电脑程序？

电脑程序即为处理输入获得输出的方式

输入 ➡️ **电脑程序** ➡️ 输出

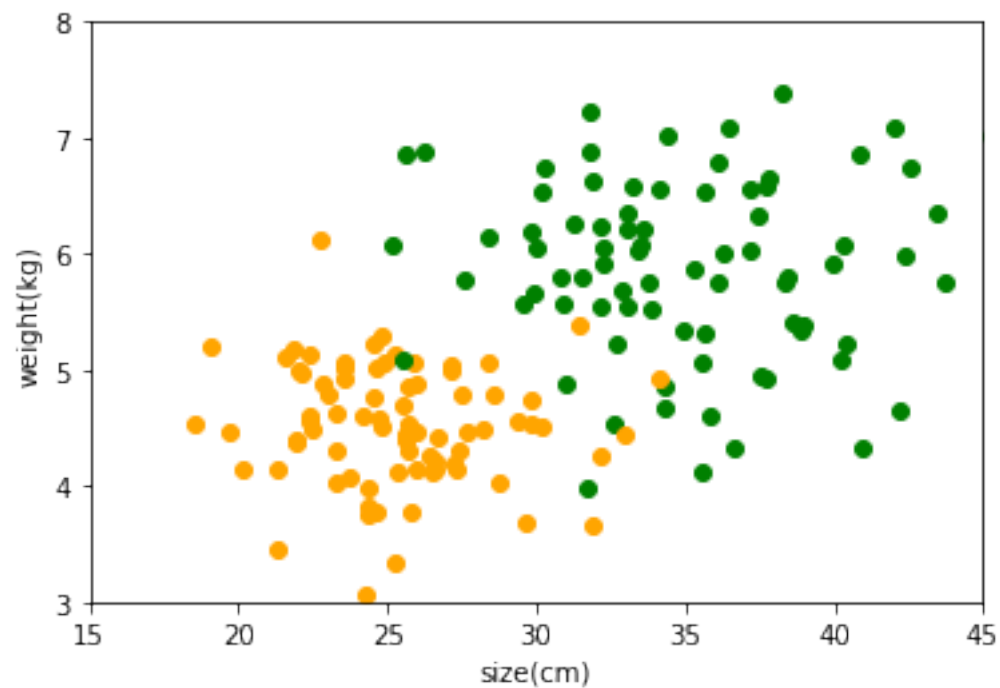需要知道从输入产生输出的规则

隐式的规则
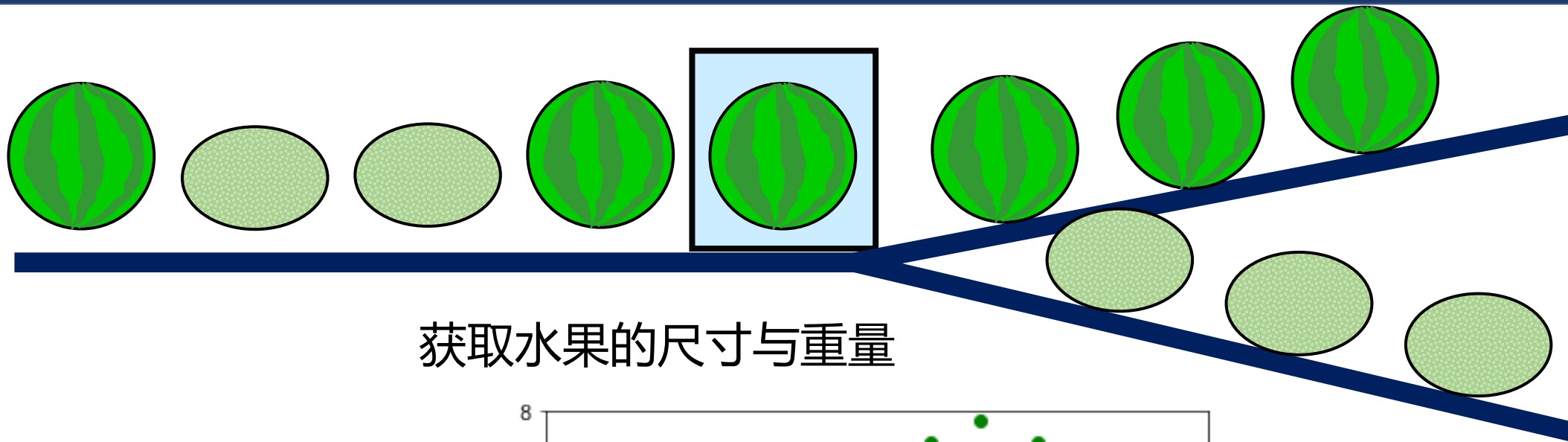
获取水果的尺寸与重量
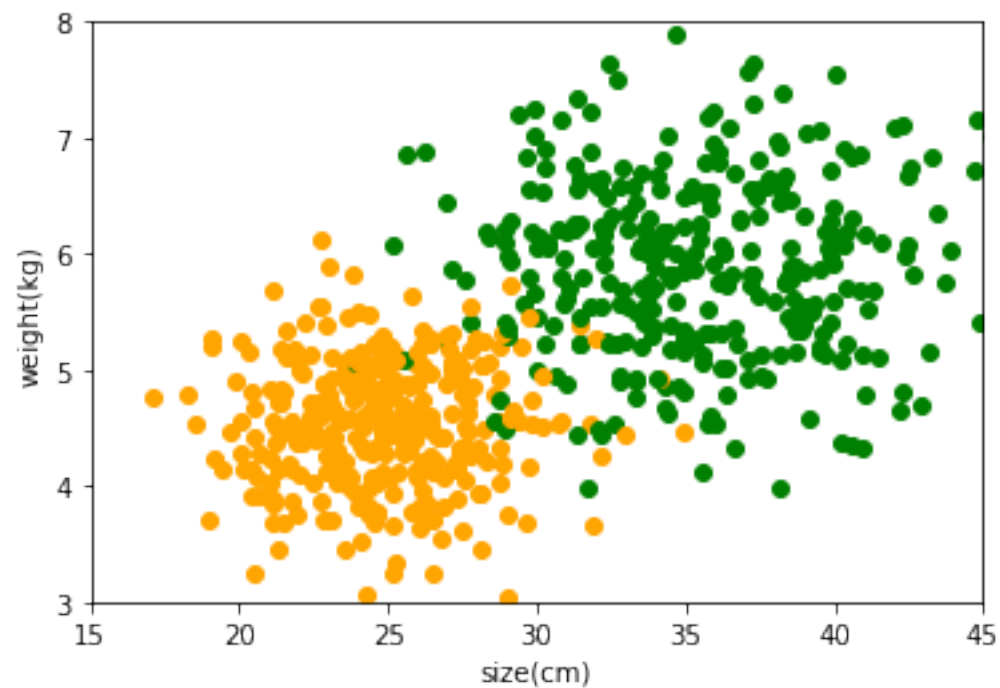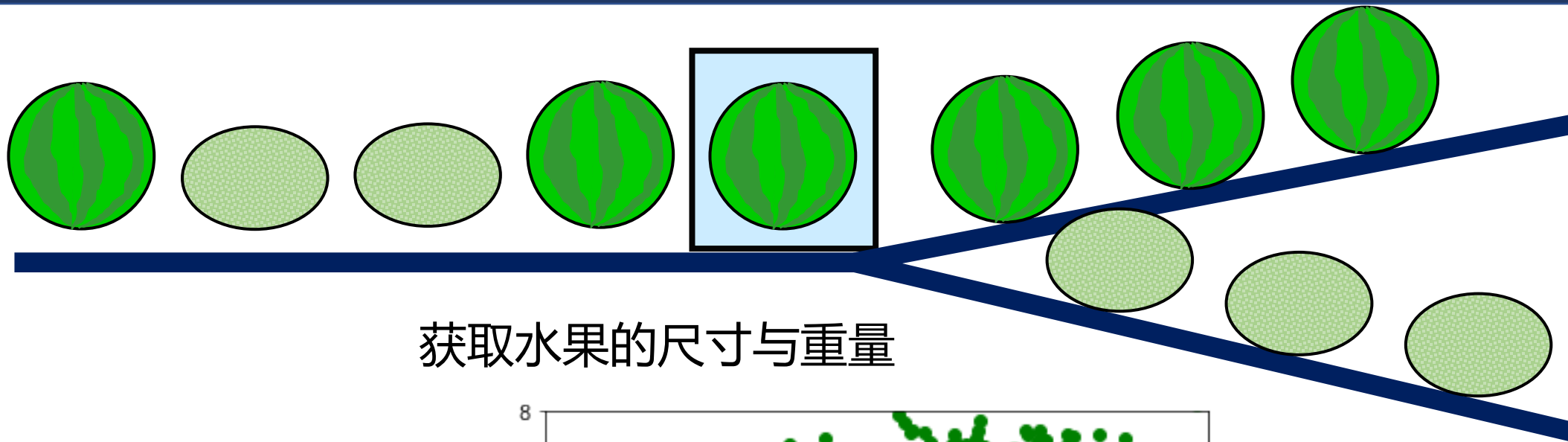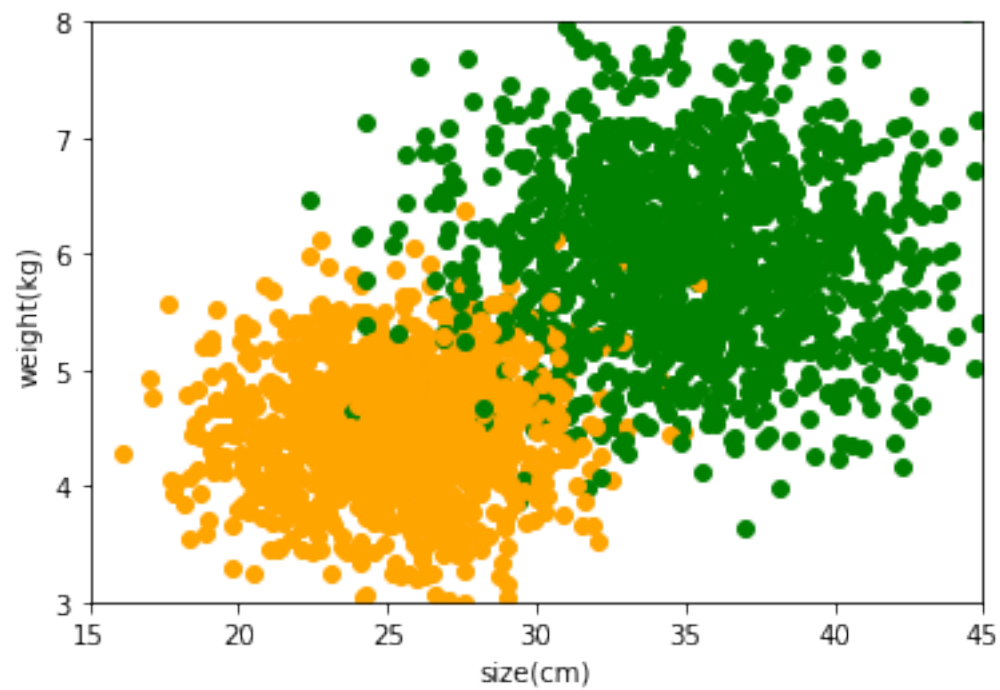
西瓜与哈密瓜分类器

获取水果的尺寸与重量

- 西瓜
- 哈密瓜

西瓜与哈密瓜分类器

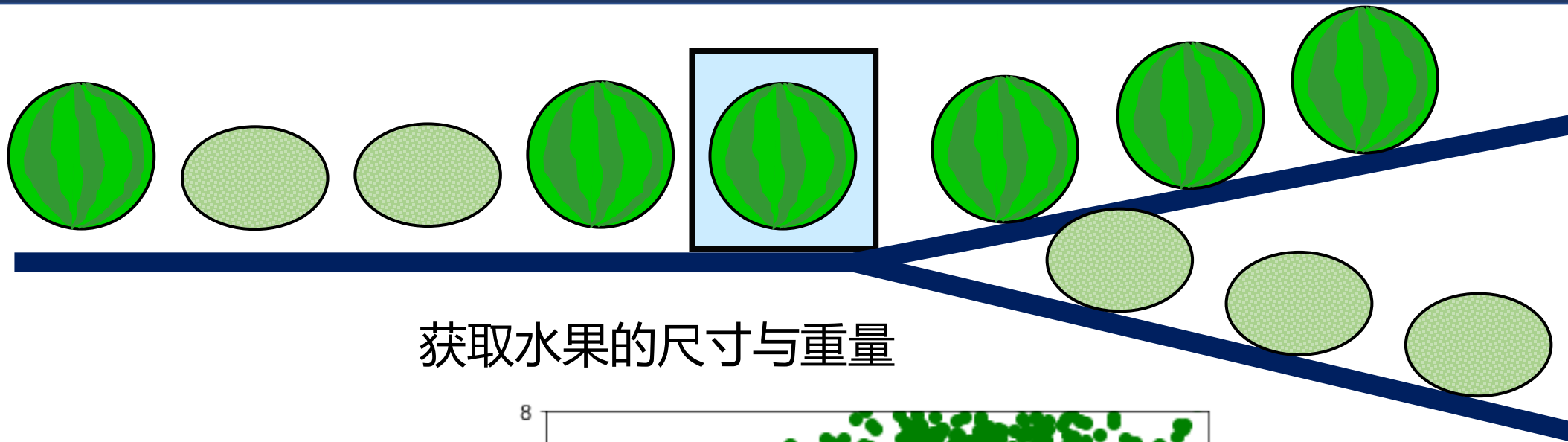获取水果的尺寸与重量

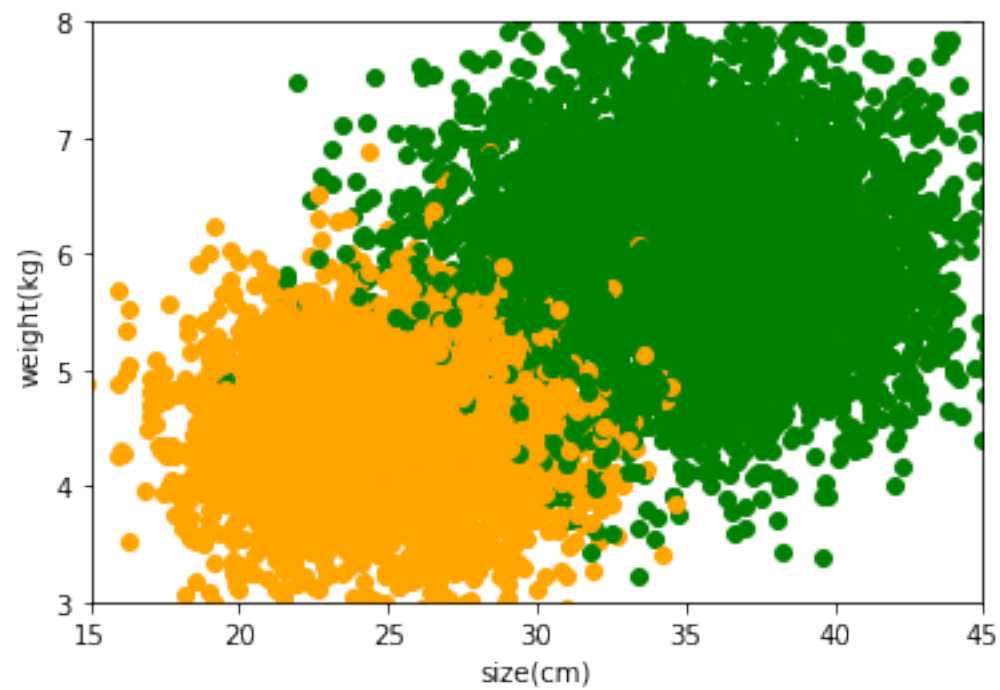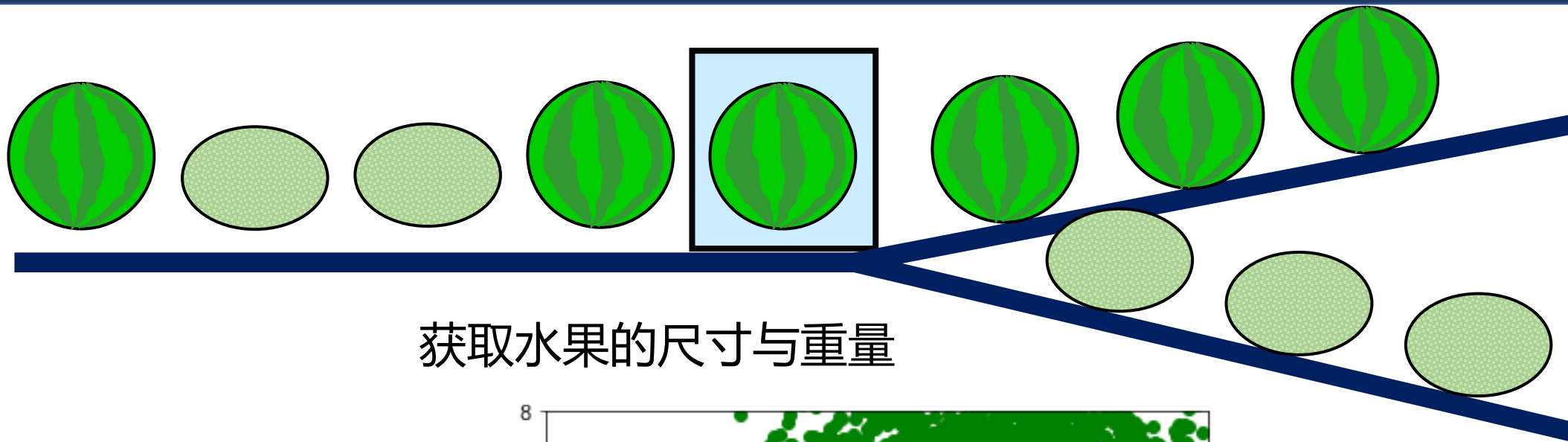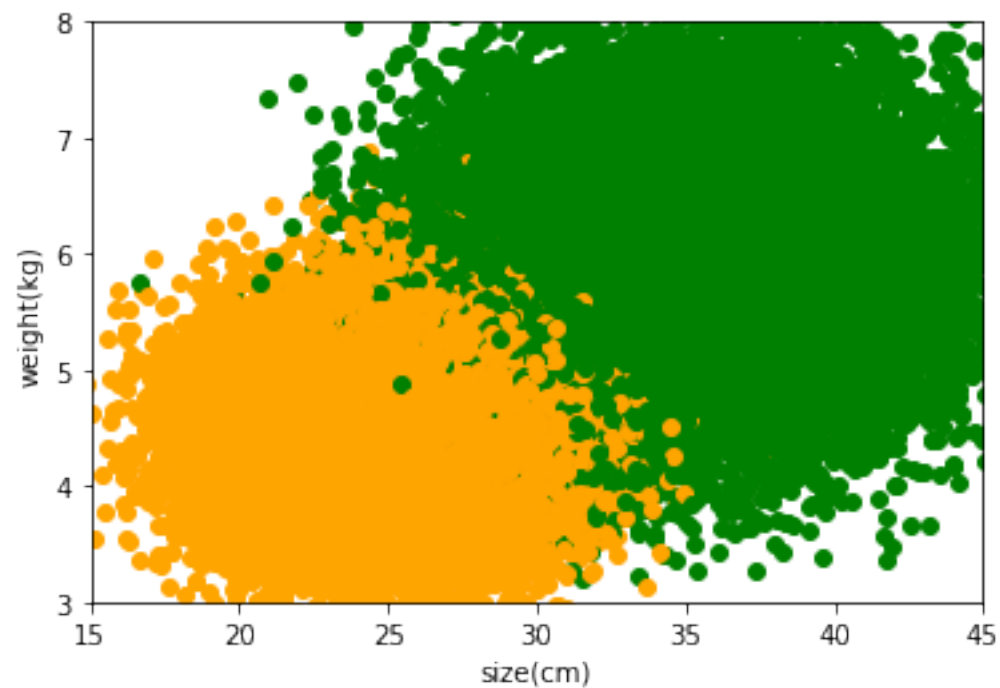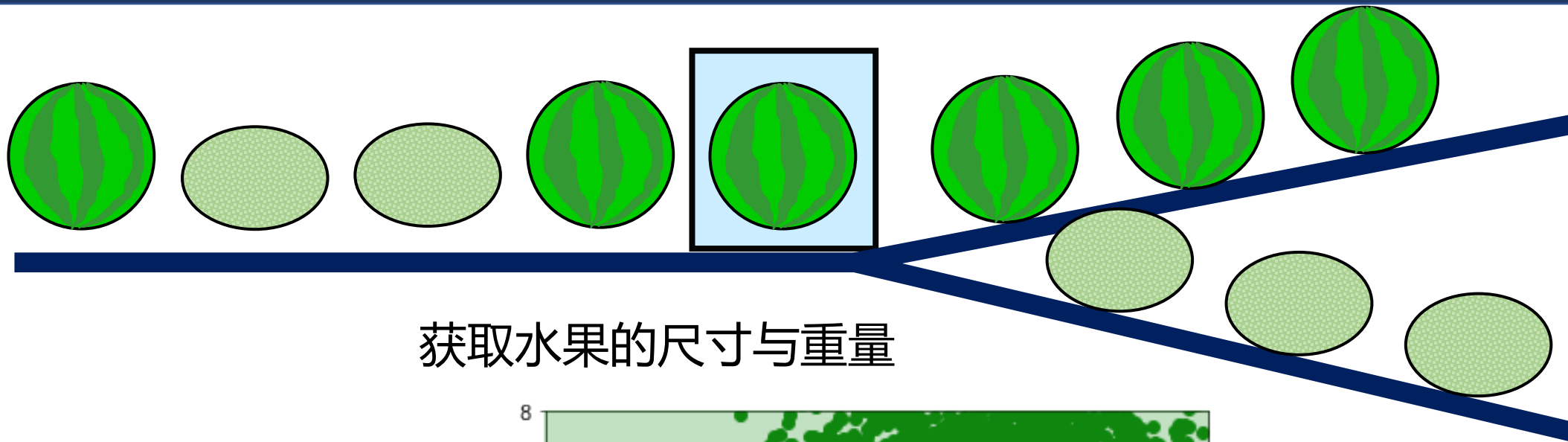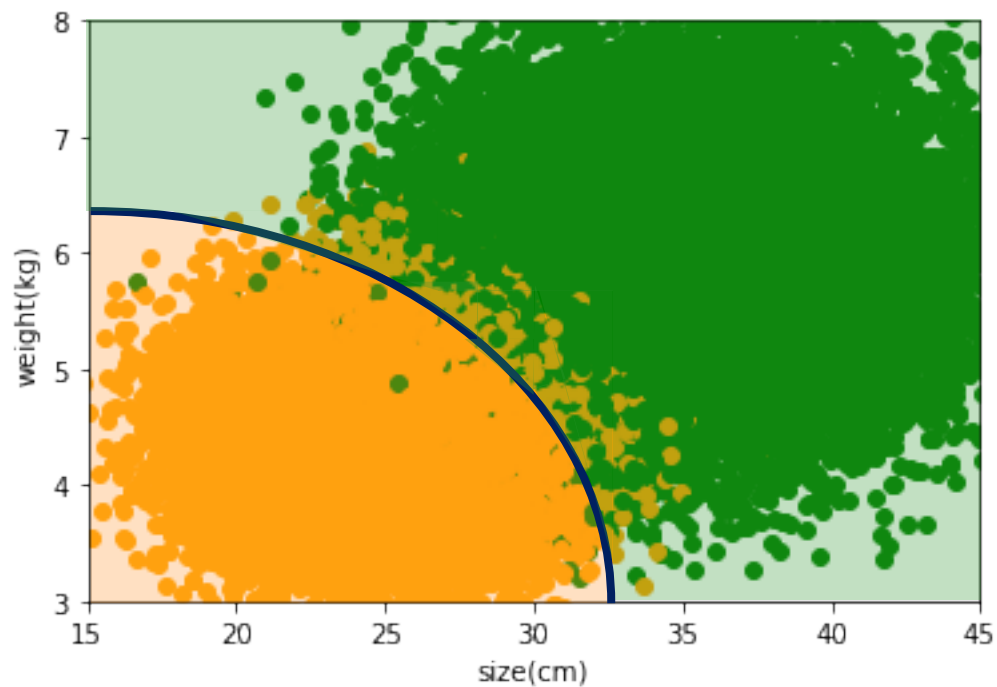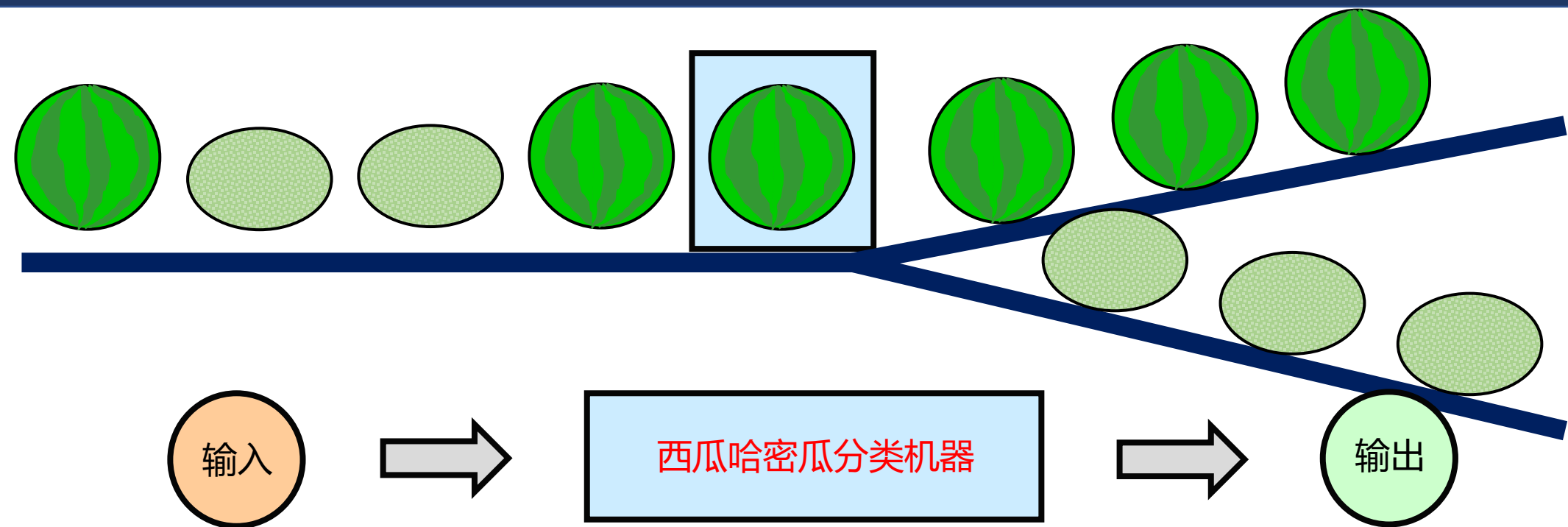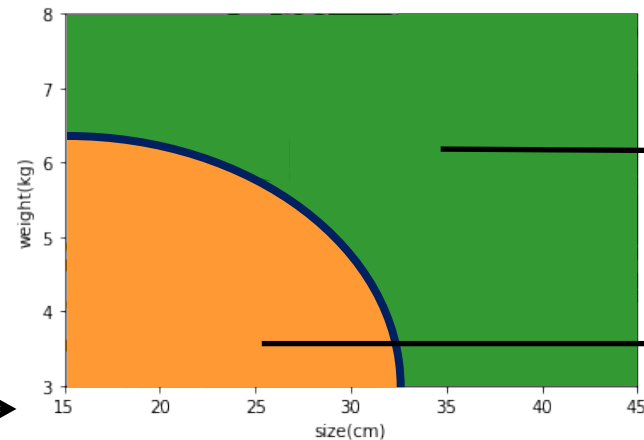- **西瓜**
- **哈密瓜**

西瓜与哈密瓜分类器

获取水果的尺寸与重量

- **西瓜**
- **哈密瓜**

西瓜与哈密瓜分类器

获取水果的尺寸与重量

- **西瓜**
- **哈密瓜**

获取水果的尺寸与重量

- **西瓜**
- **哈密瓜**

西瓜与哈密瓜分类器

获取水果的尺寸与重量

- **西瓜**
- **哈密瓜**

西瓜与哈密瓜分类器

获取水果的尺寸与重量

- **西瓜**
- **哈密瓜**

获取水果的尺寸与重量

- **西瓜**
- **哈密瓜**

输入

西瓜哈密瓜分类机器

输出

重量

尺寸

西瓜或哈密瓜

特征、描述符

类标签

输入

x

机器学习模型

输出

y

输入

x_1, x_2

机器学习模型

输出

y

输入　　　　　　　　　　　　　　　　　　　　　　　　输出

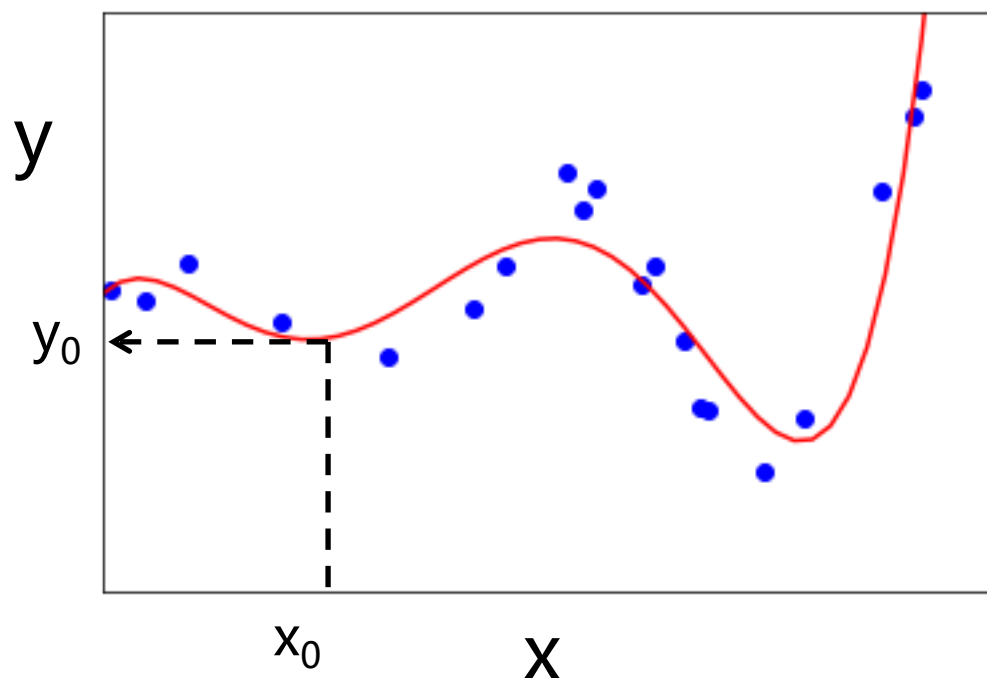$x_1, x_2, ..., x_n$ ⟶ | 机器学习模型 | ⟶ y

6 ⟶ 6
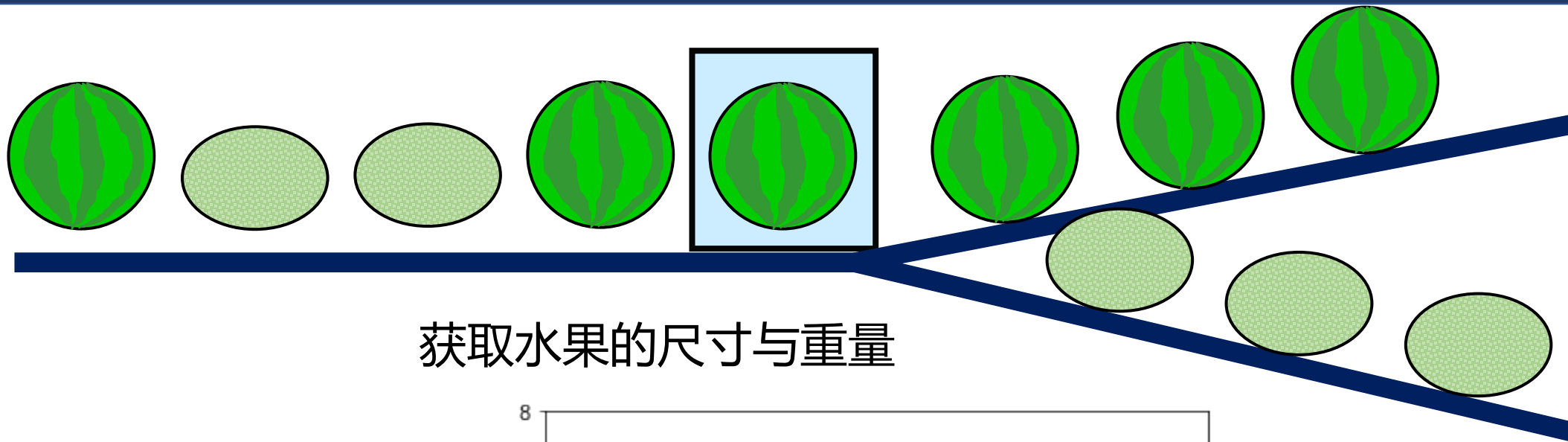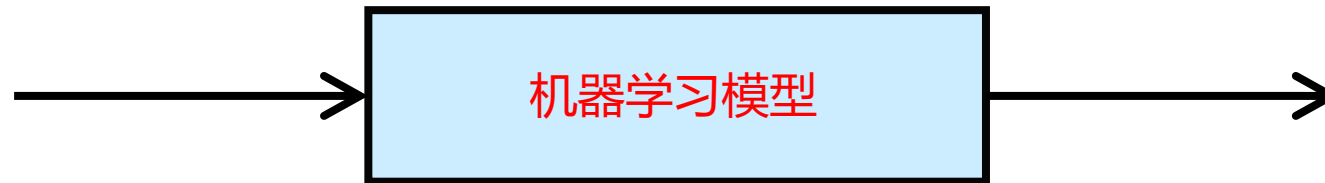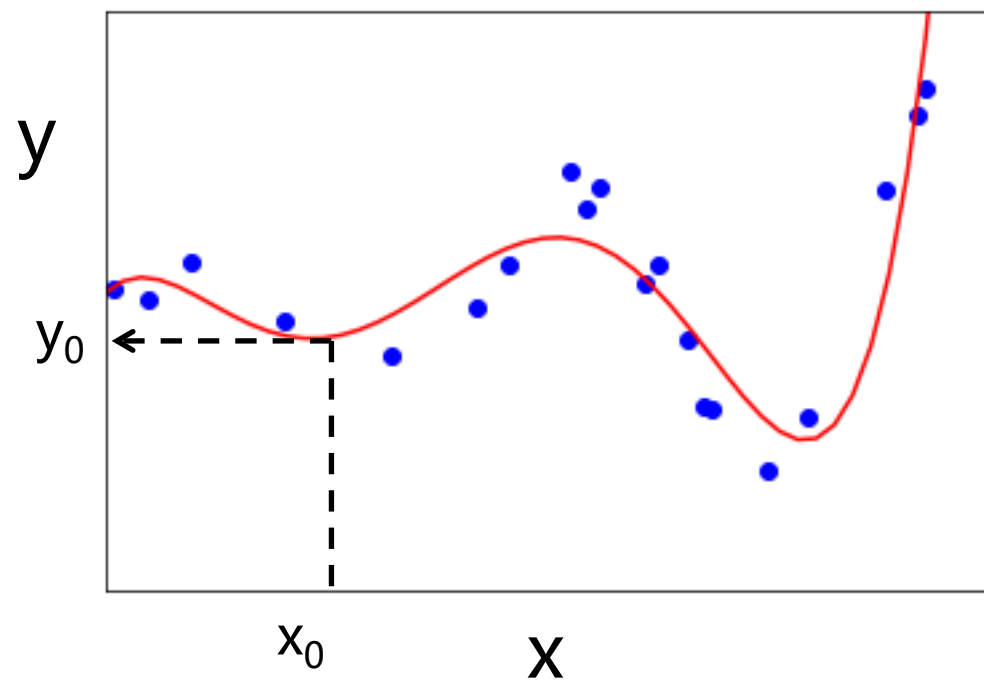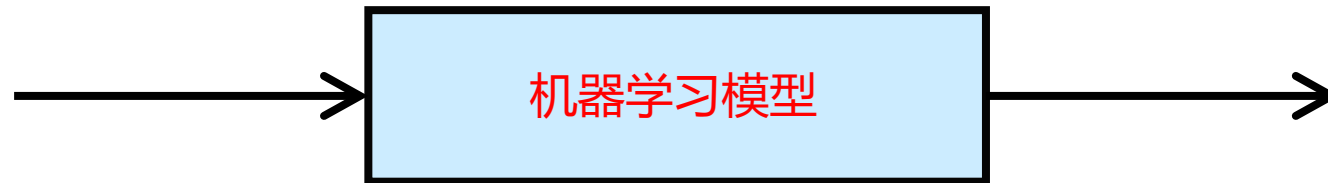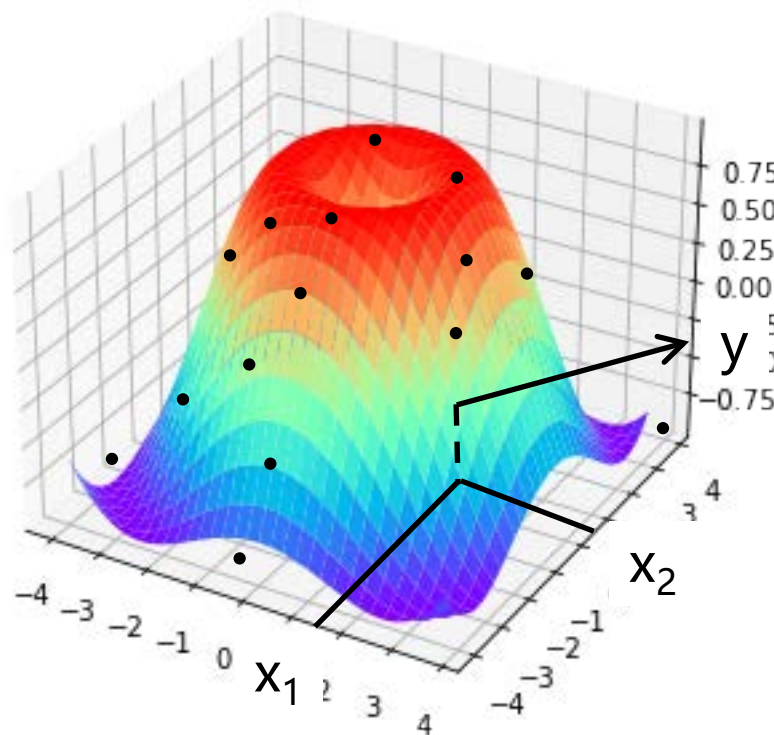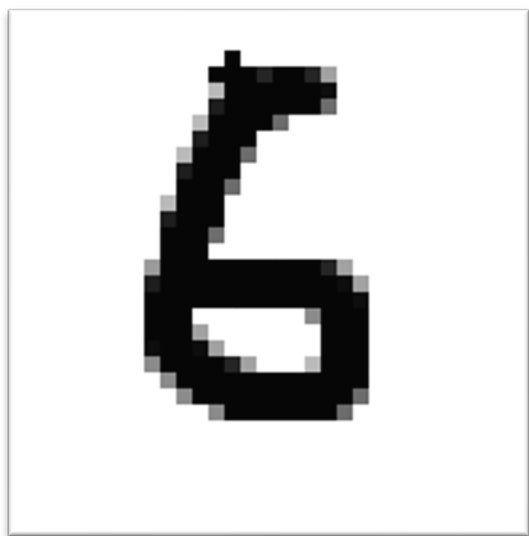
寻找高维到低维的映射关系

32×32

机器学习有价值的地方在于通过已有的数据对未知的数据进行预测
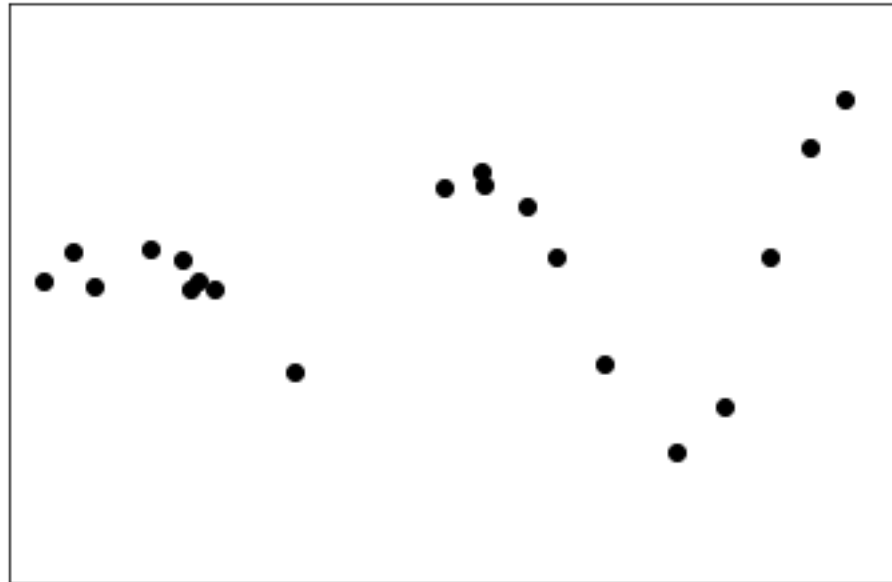
数据　　　模型　　　预测
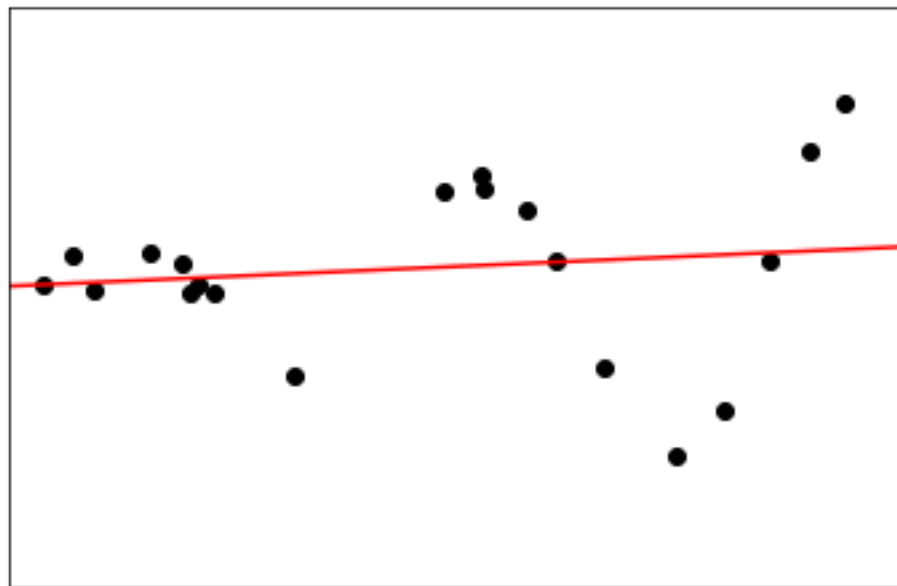
通过函数拟合对原有数据进行插值

魔法?

机器像人一样的思考?

函数拟合插值

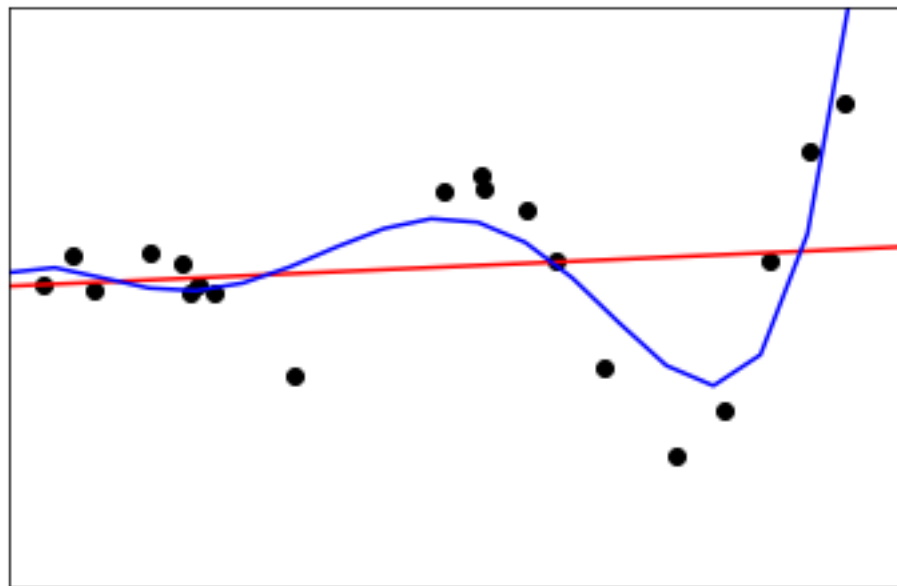但是，高维函数的拟合并没有想象的简单

为什么机器学习比想象的复杂?

以曲线拟合为例

为什么机器学习比想象的复杂？

以曲线拟合为例

模型1

为什么机器学习比想象的复杂?

以曲线拟合为例

模型2

模型1

为什么机器学习比想象的复杂?

以曲线拟合为例

寻找映射关系
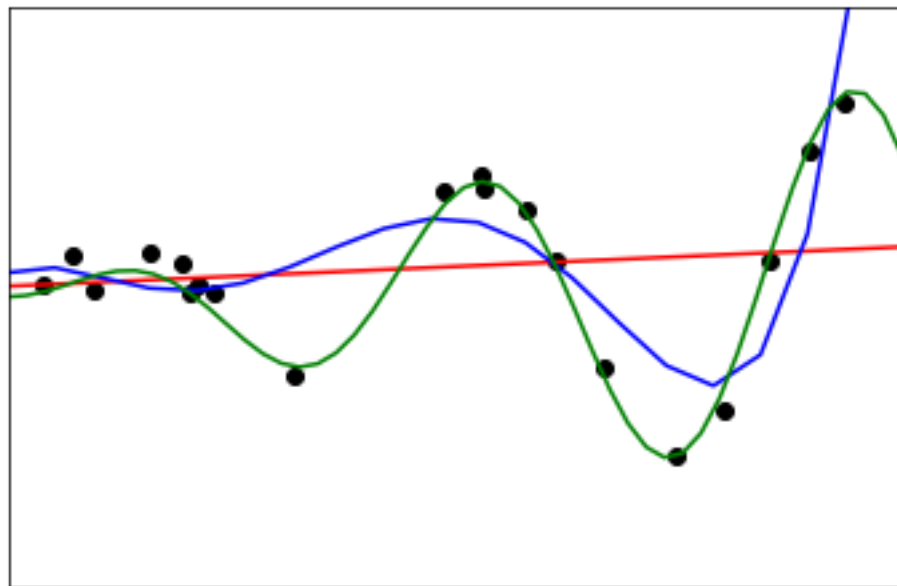
模型2

模型3

模型1

多种拟合方式，并不知道哪种是最好的

需要选择最合适的模型

目录

1. 机器学习是什么

2. 机器学习与化学研究

3. 机器学习库

Science is changing, the tools of science are changing. And that requires different approaches.

---- Erich Bloch, 1925-2016

斯坦福大学2021年AI数据生成索引报告：

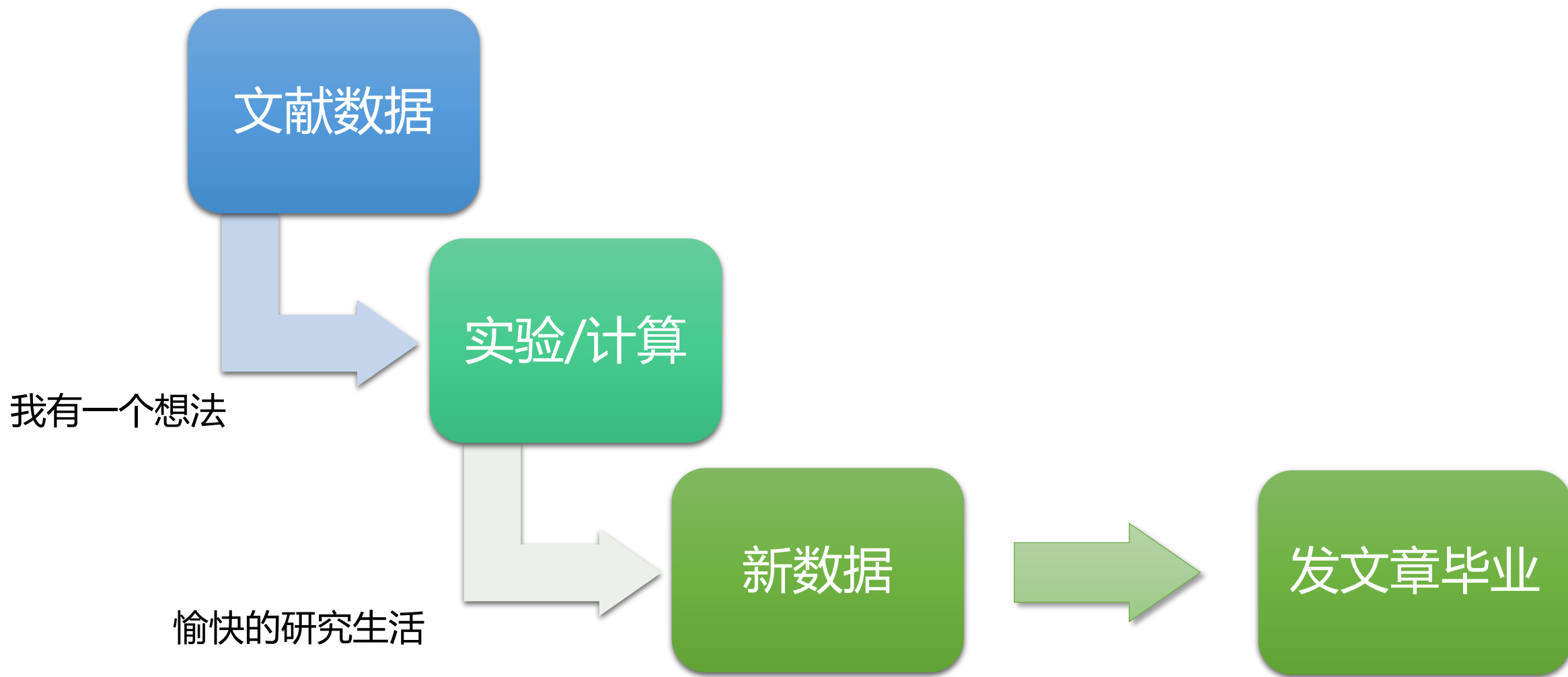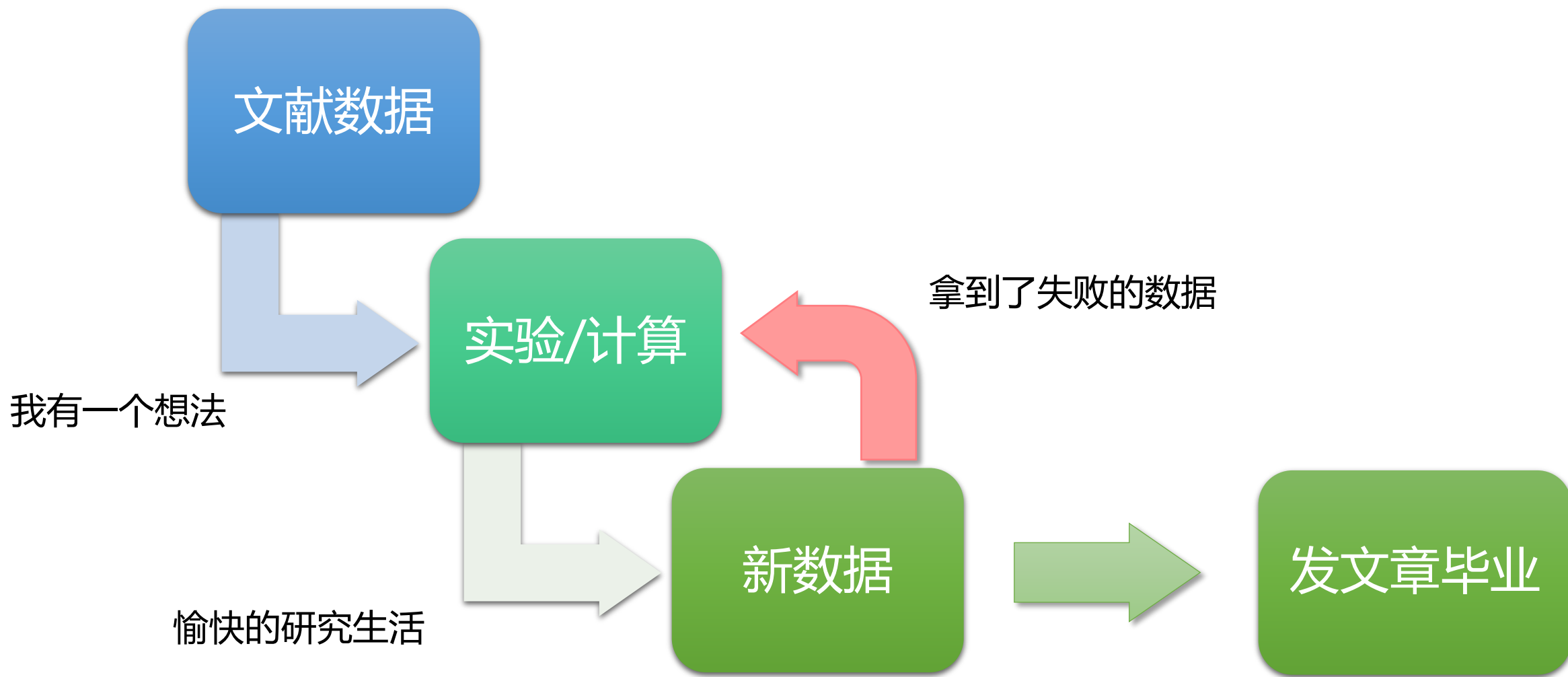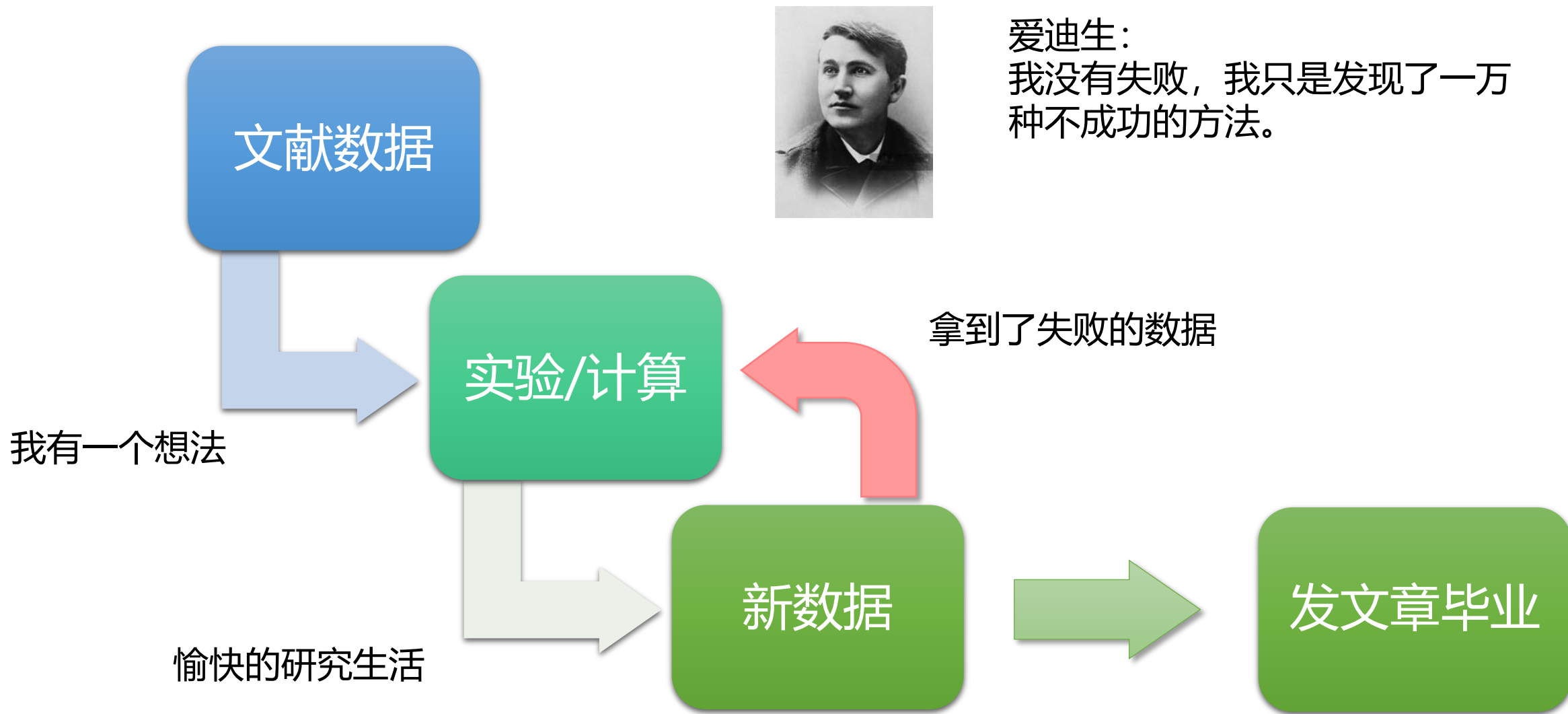　　涉及AI的期刊出版物数量于2019年到2020年之间增长了**34.5%**，而在2018年到2019年此增长率仅为**19.6%**。

　　2019年，人工智能类出版物占全球所有经同行评审科学类出版物的**3.8%**，相较2011年的**1.3%**有所提升。

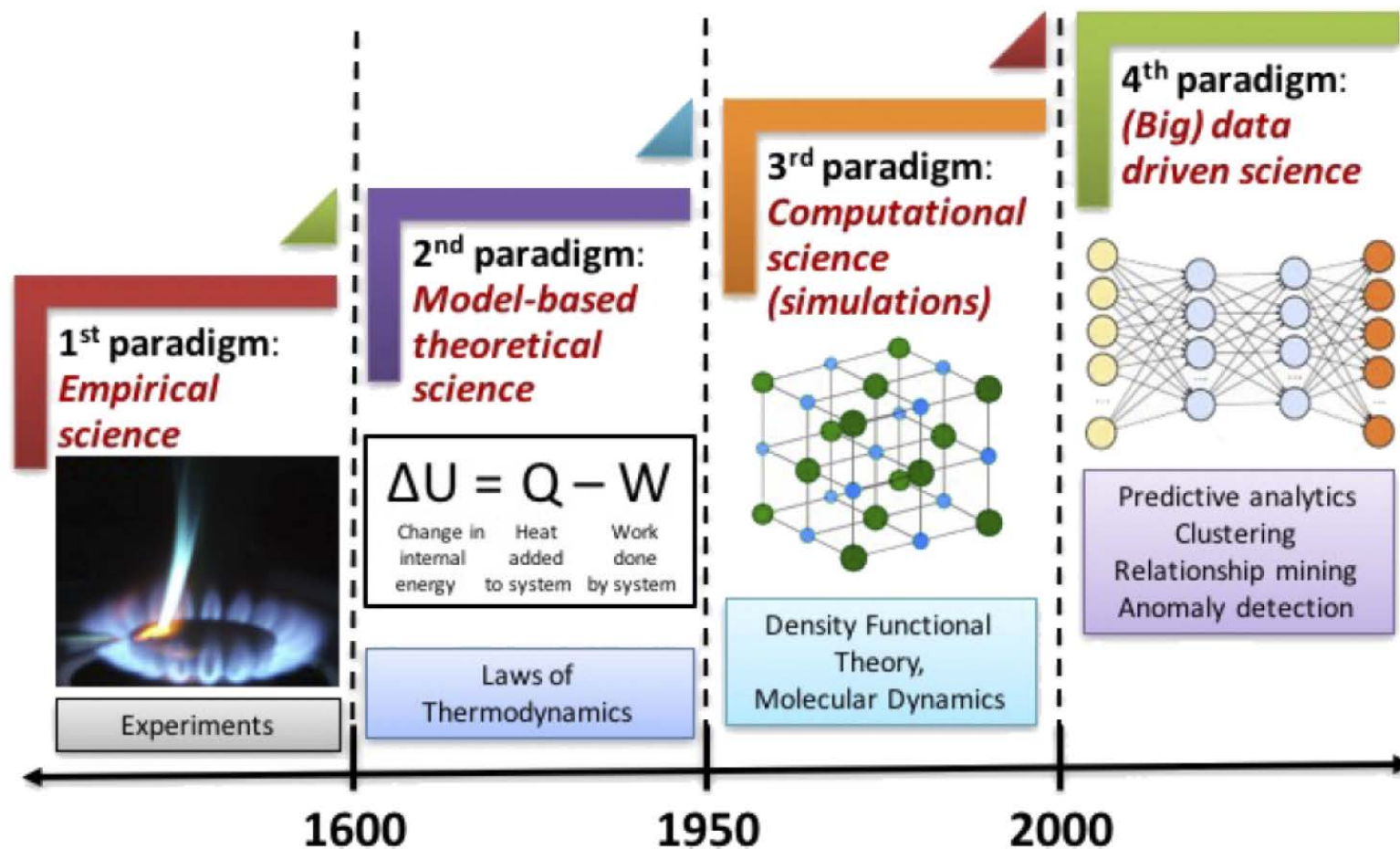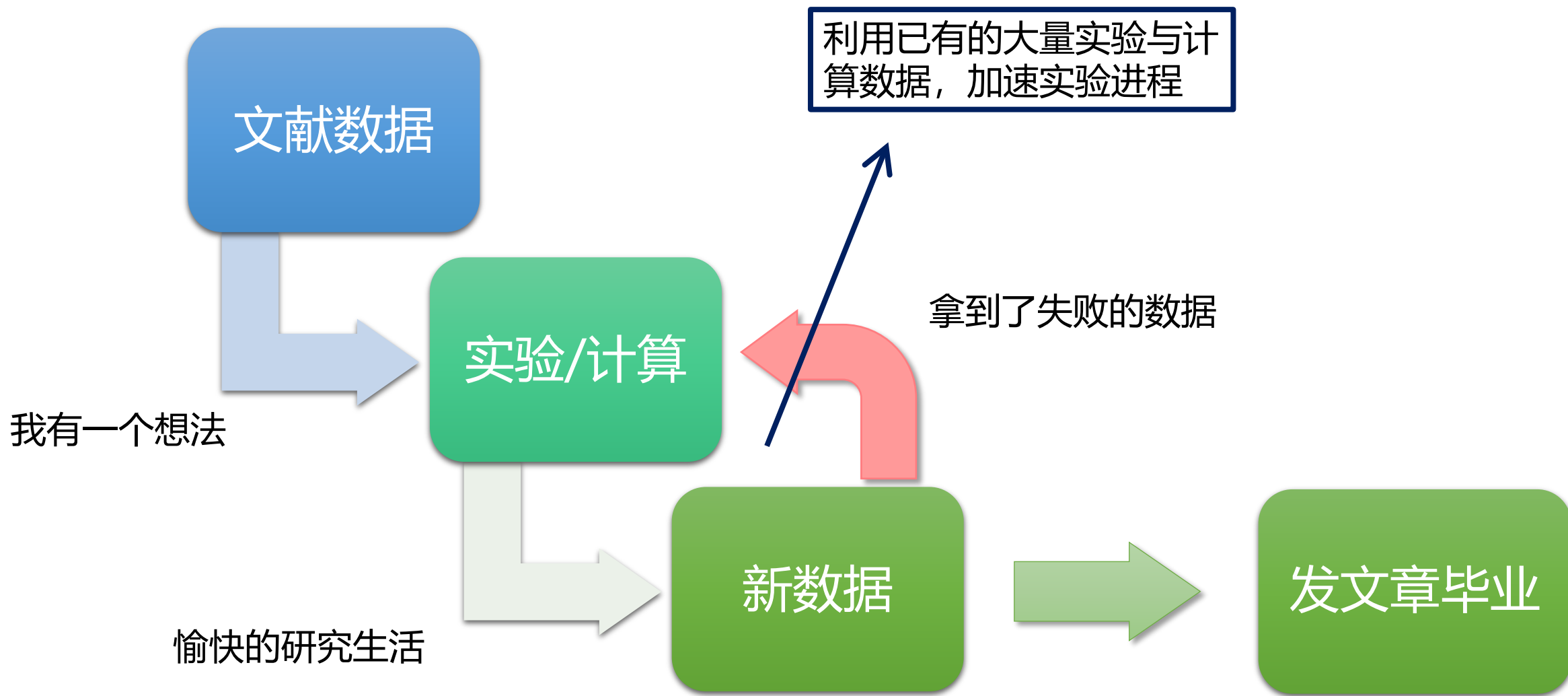G. R. Schleder, et. al. *J. Phys. Mater.* **2019**, 2, 032001.

文献数据

实验/计算

我有一个想法

新数据

发文章毕业

愉快的研究生活

文献数据

实验/计算

拿到了失败的数据

我有一个想法

愉快的研究生活

新数据

发文章毕业

文献数据

爱迪生：
我没有失败，我只是发现了一万种不成功的方法。

实验/计算

拿到了失败的数据

我有一个想法

愉快的研究生活

新数据

发文章毕业

但是，现在已经是2022年了。。。

A. Agrawal and A. Choudhary, APL Mater. 2016, 4, 053208.

文献数据

利用已有的大量实验与计算数据，加速实验进程

实验/计算

拿到了失败的数据

我有一个想法

愉快的研究生活

新数据

发文章毕业

目录

1. 机器学习是什么

2. 机器学习与化学研究

3. 机器学习库

Python机器学习库

	scikit-learn	mlpy	PyBrain	Keras
名称	scikit-learn	mlpy	PyBrain	Keras
简介	大名鼎鼎，基于 NumPy, SciPy, Matplotlib的开源机器学习工具包	同样建立在 NumPy/SciPy 和 GNU 科学库之上	为机器学习任务提供灵活、易应、强大的机器学习算法。	开源人工神经网络库，可以作为Tensorflow、Theano等高阶应用的程序接口
功能	分类，回归和聚类算法	分类，回归和聚类等	神经网络 强化学习 进化算法	神经网络

scikit-learn库

https://scikit-learn.org