

声明：本课程版权归华算科技所有，仅限个人学习，严禁任何形式的录制、传播和账号分享。一经发现，平台将依法保留追究权，情节严重者将承担法律责任。

# 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



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

1. 机器学习是什么
2. 机器学习与化学研究
3. 机器学习库

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



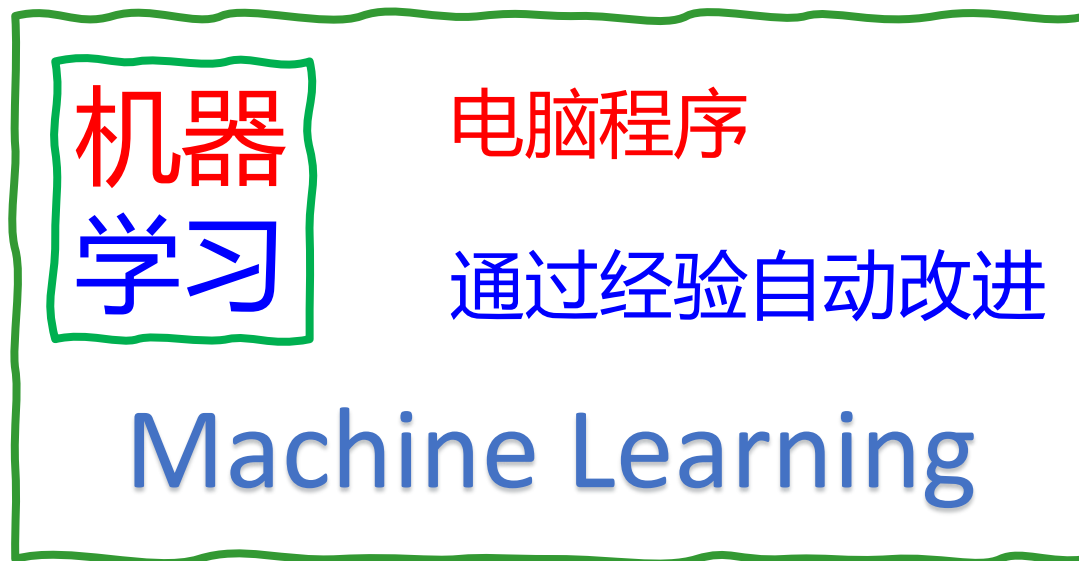
# 什么是机器学习?



# 什么是机器学习?

Concerned with the question of how to construct **computer programs** that automatically improve with experience.

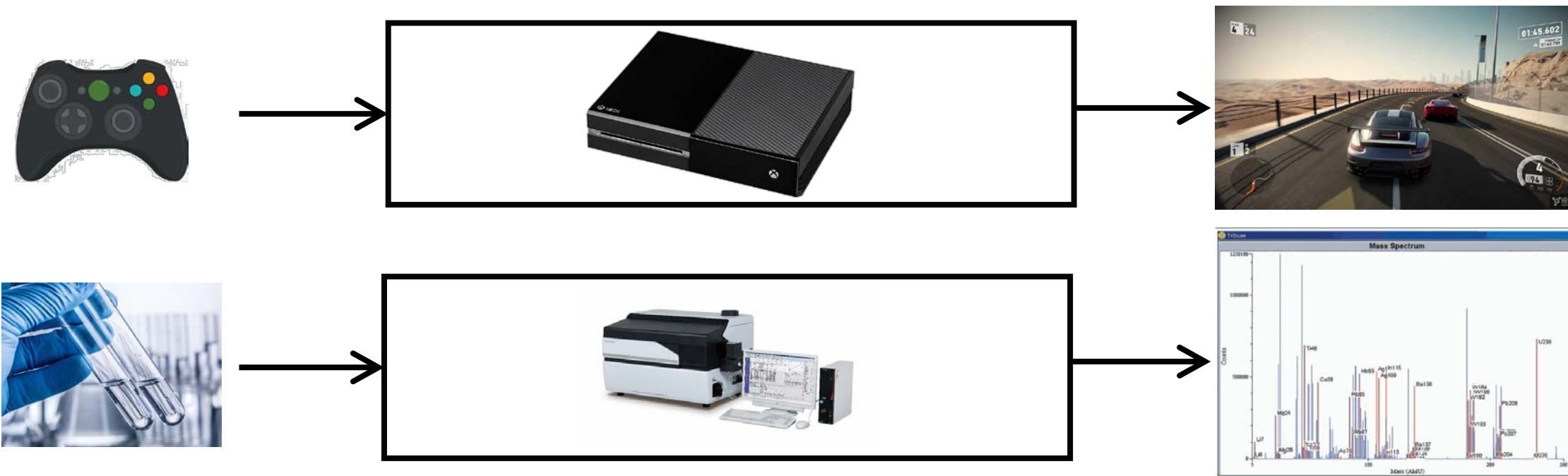
——Tom Mitchell





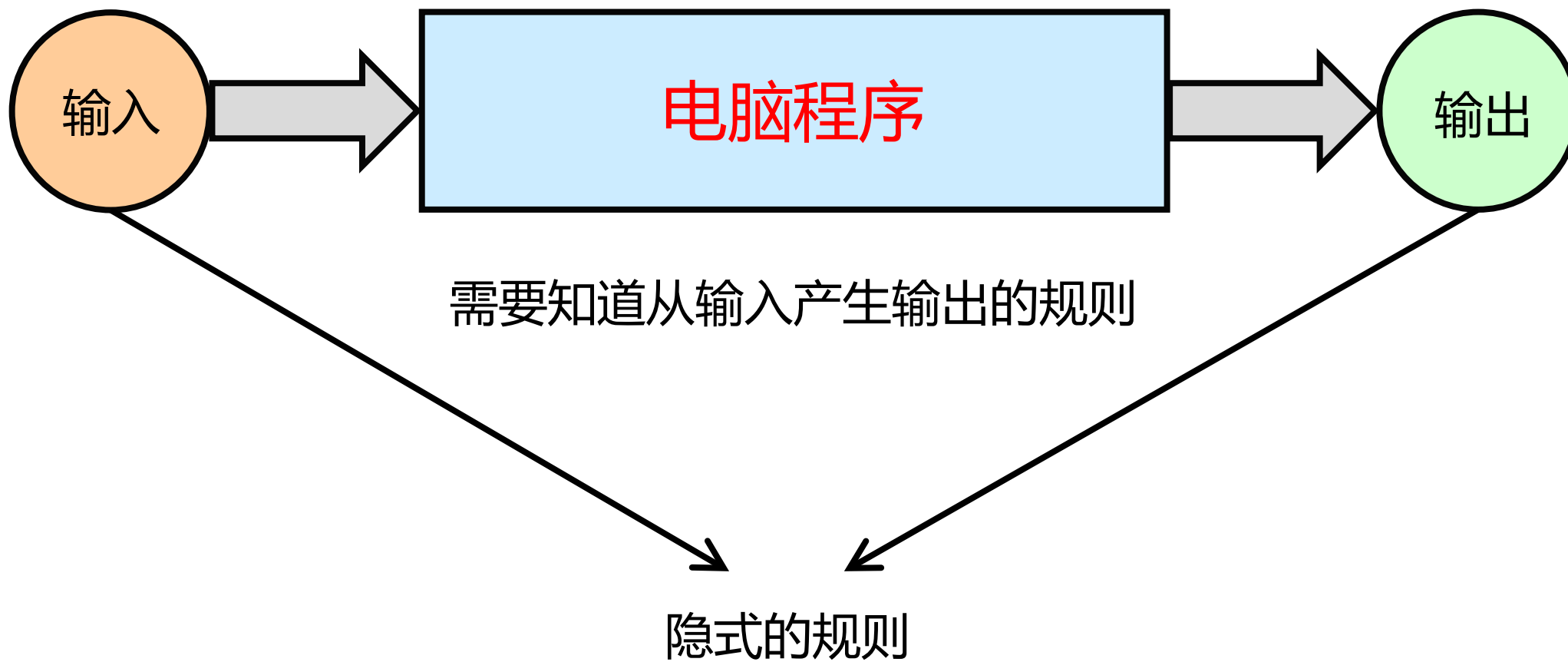
# 电脑程序？

# 电脑程序即为处理输入获得输出的方式

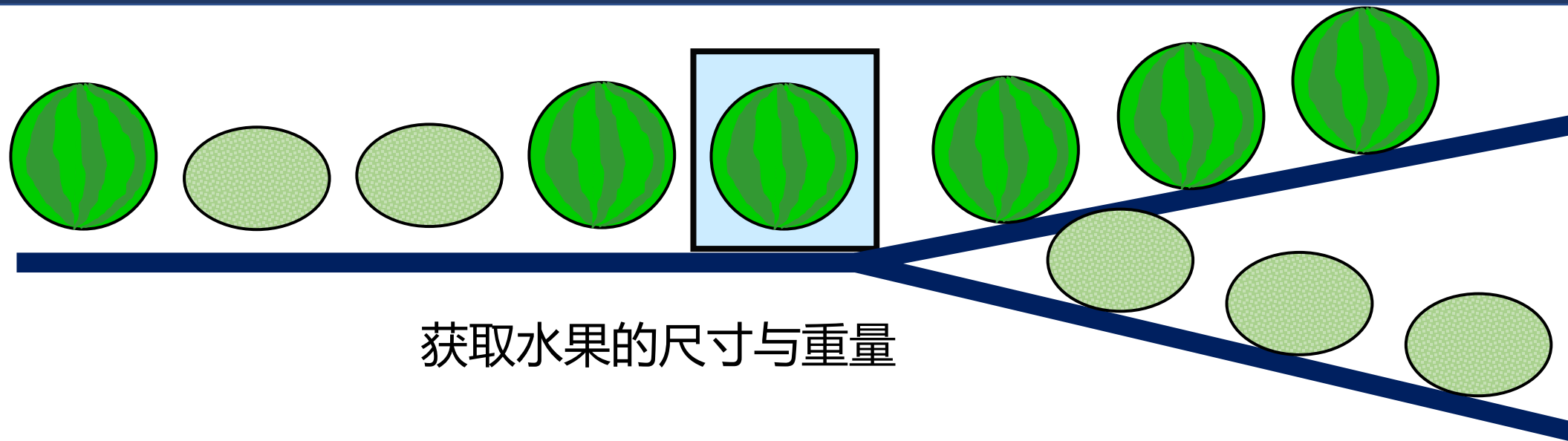


# 电脑程序？

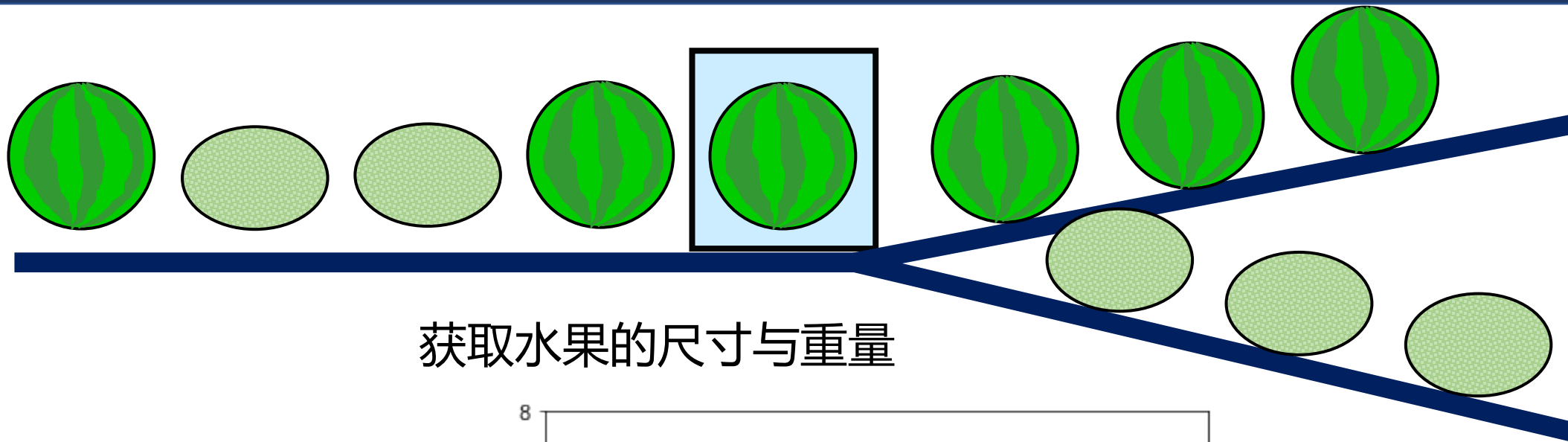
电脑程序即为处理输入获得输出的方式



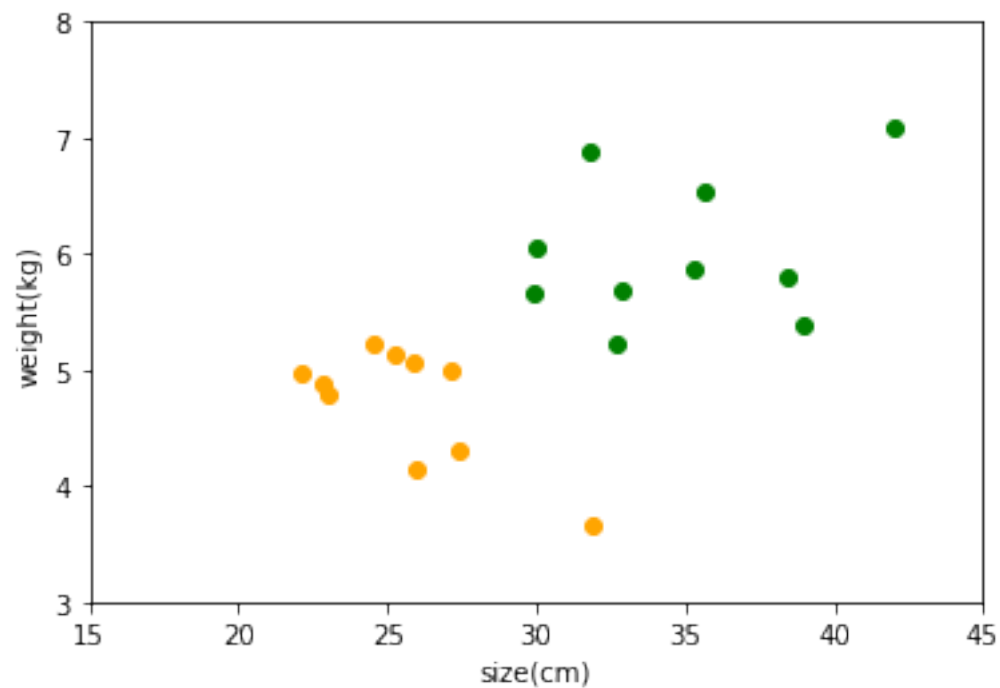
# 西瓜与哈密瓜分类器



# 西瓜与哈密瓜分类器

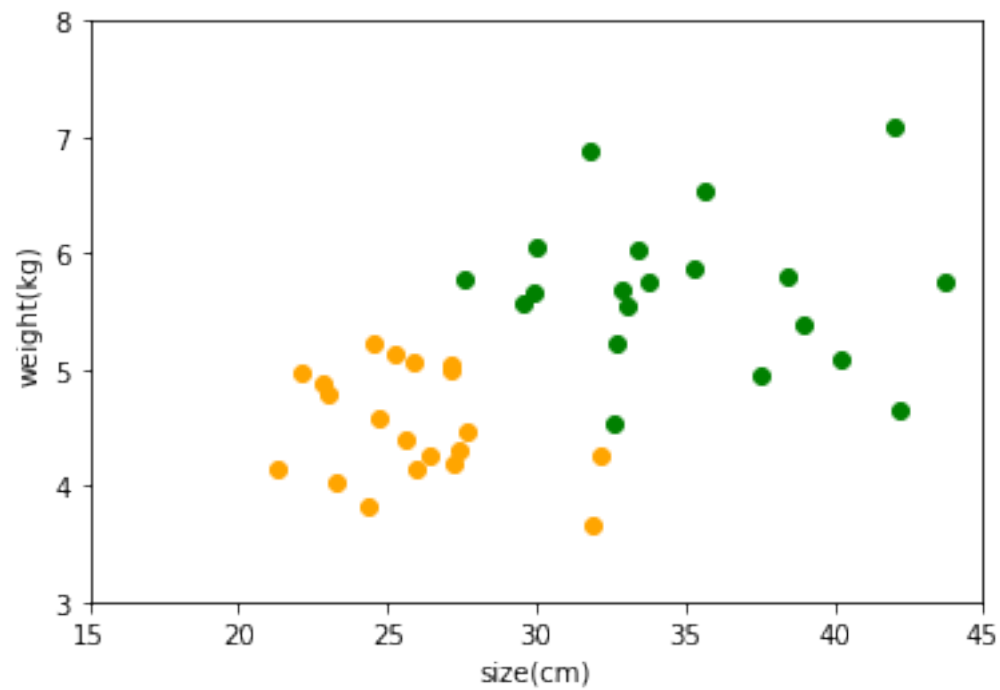
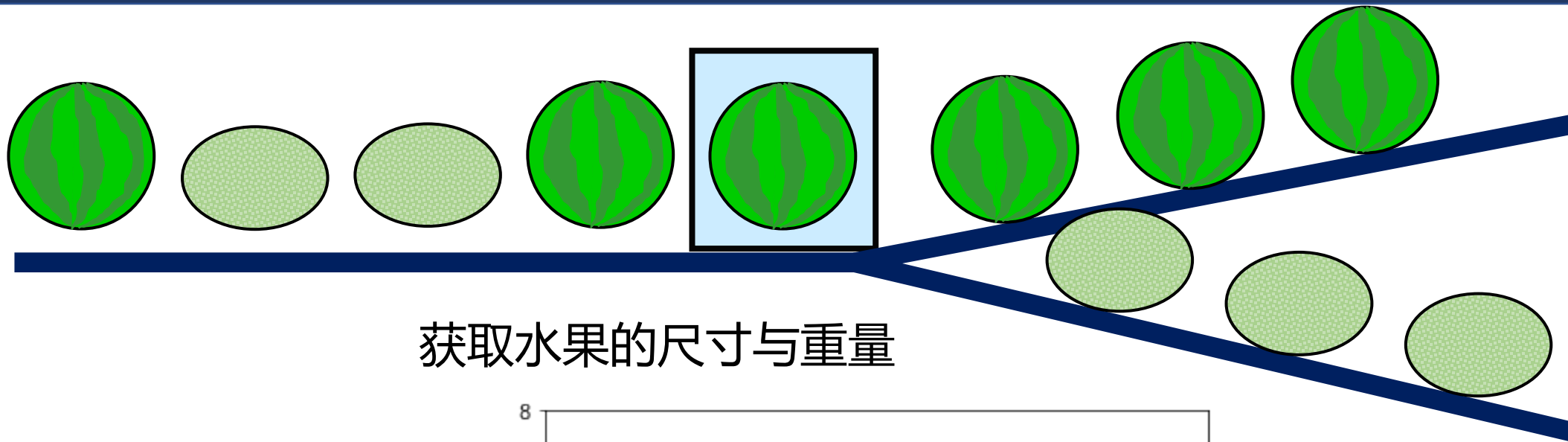


获取水果的尺寸与重量



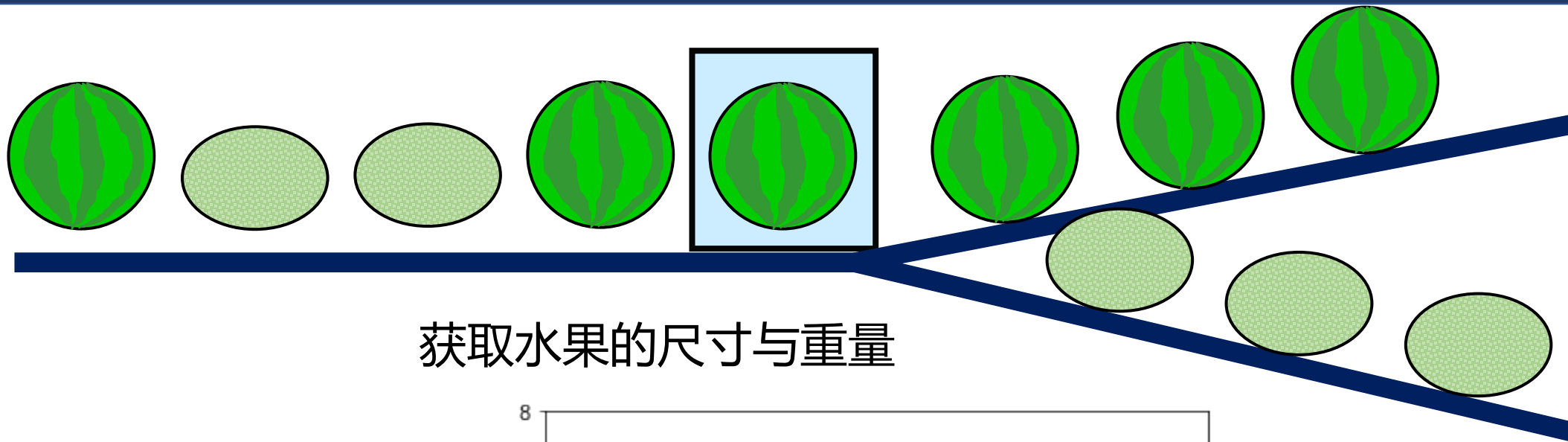
- 西瓜
- 哈密瓜

# 西瓜与哈密瓜分类器

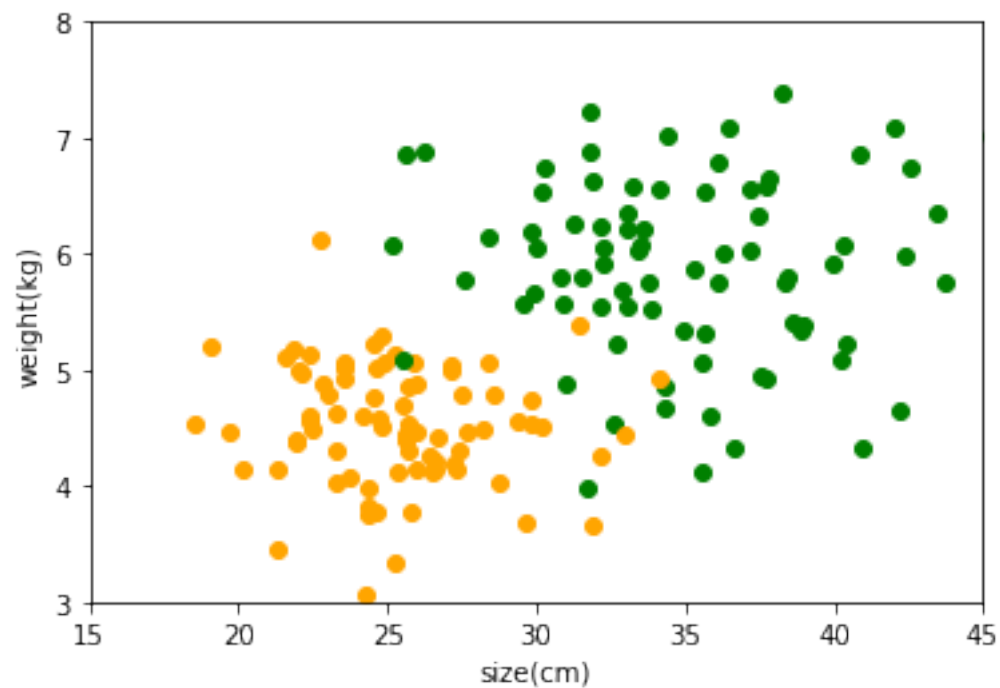


- 西瓜
- 哈密瓜

# 西瓜与哈密瓜分类器

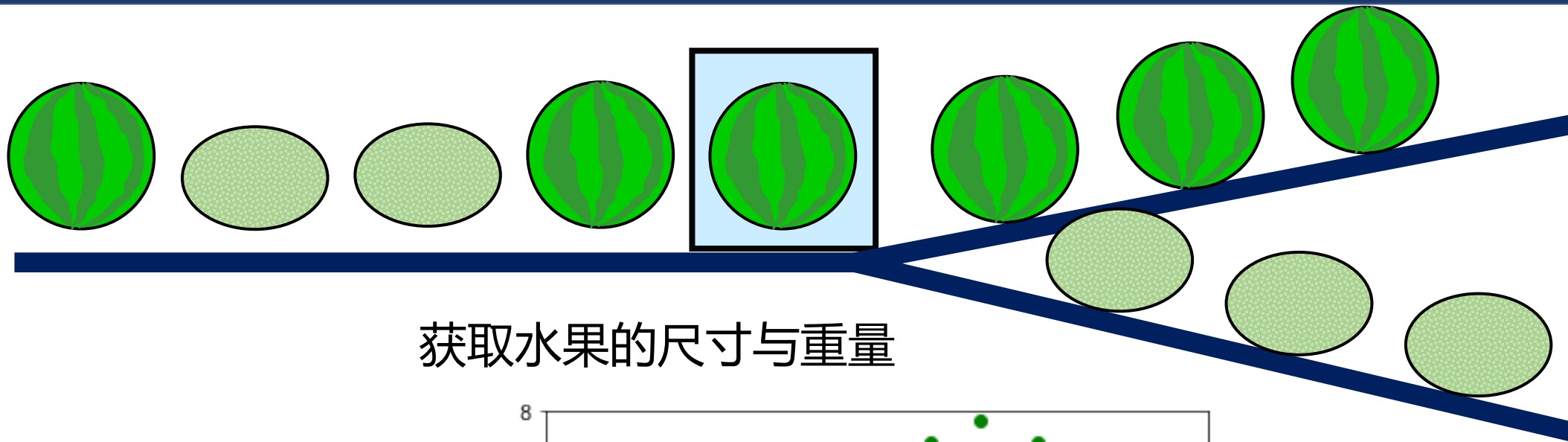


获取水果的尺寸与重量

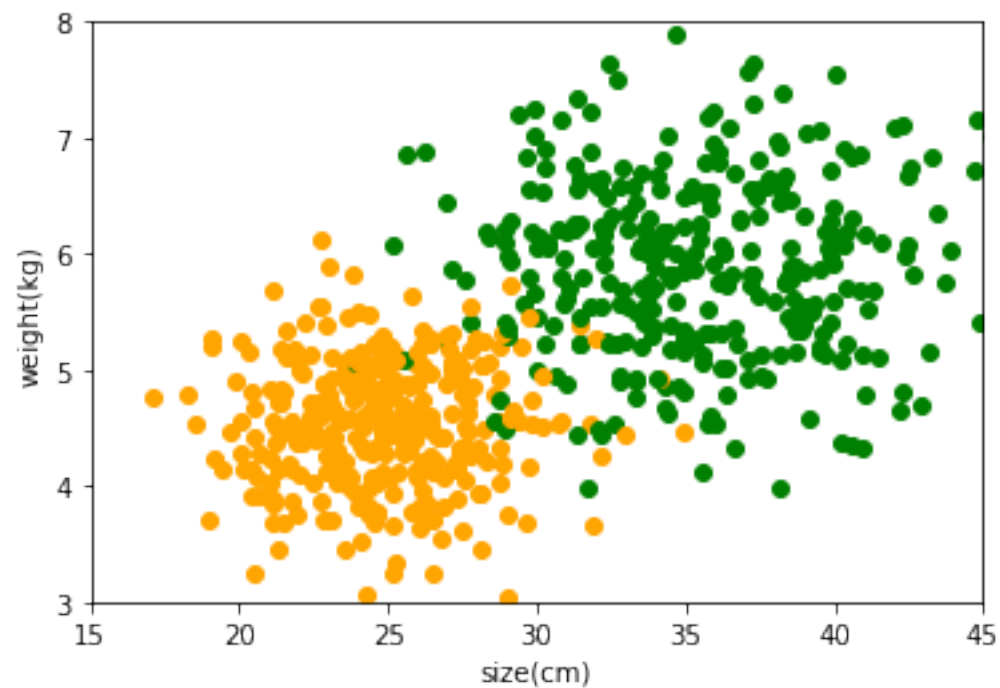


- 西瓜
- 哈密瓜

# 西瓜与哈密瓜分类器

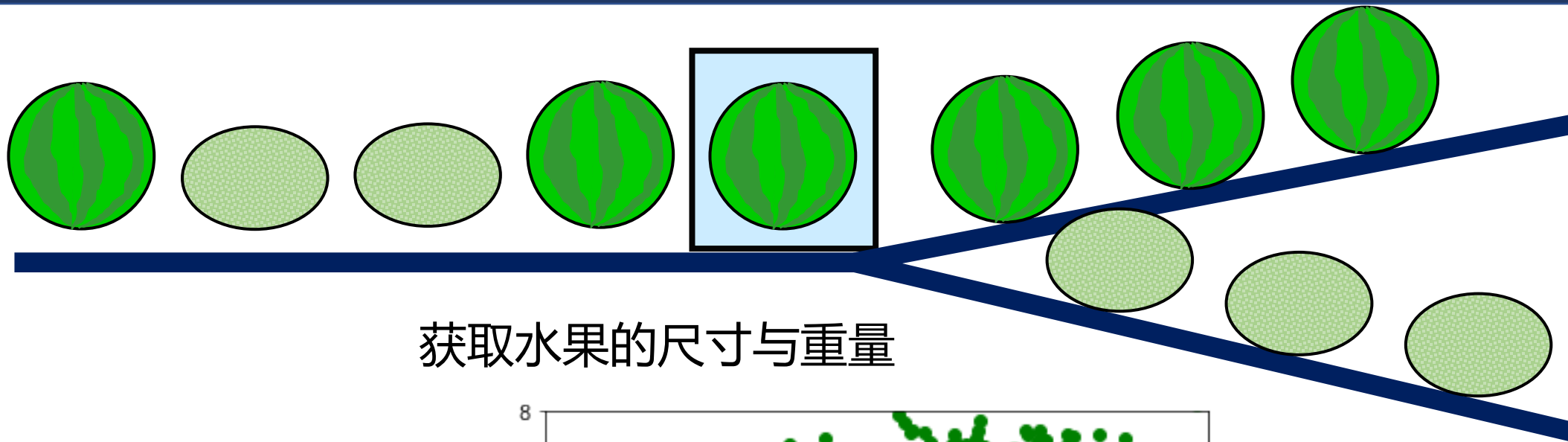


获取水果的尺寸与重量

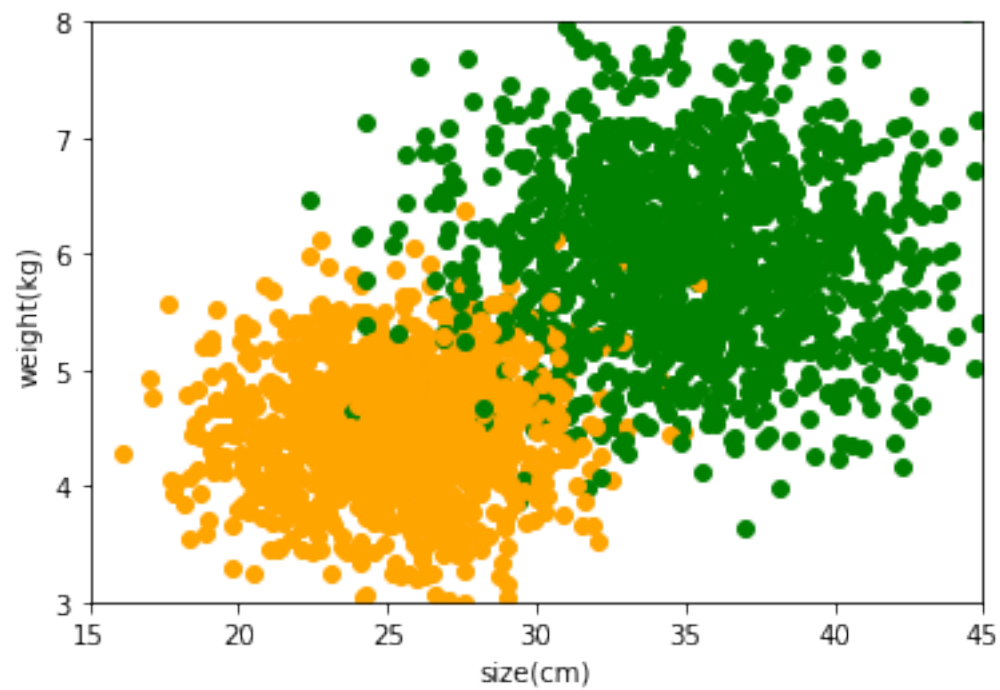


- 西瓜
- 哈密瓜

# 西瓜与哈密瓜分类器



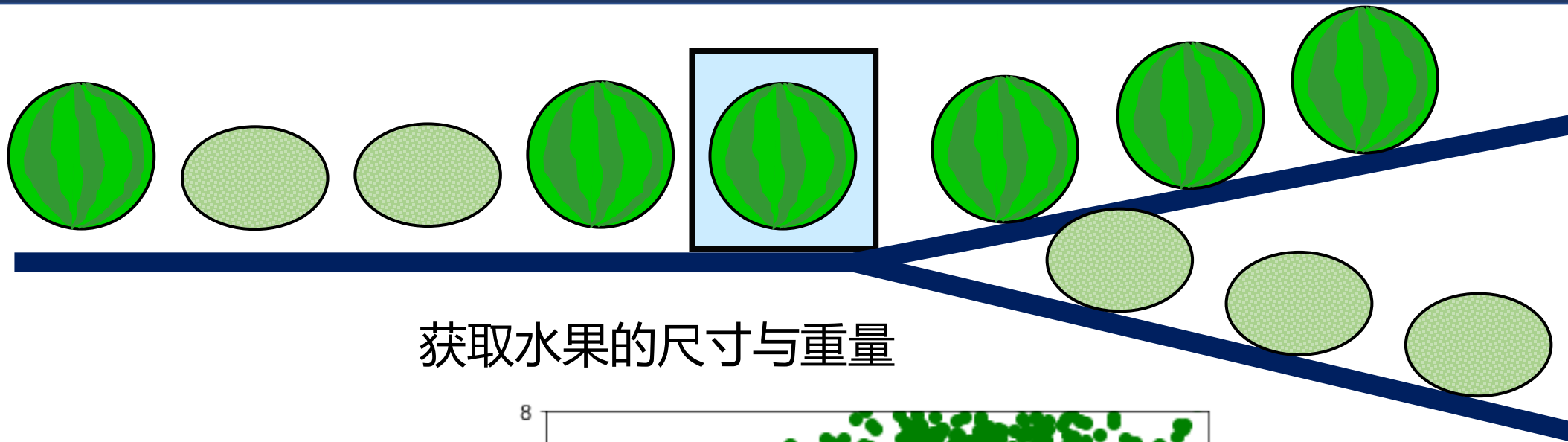
获取水果的尺寸与重量



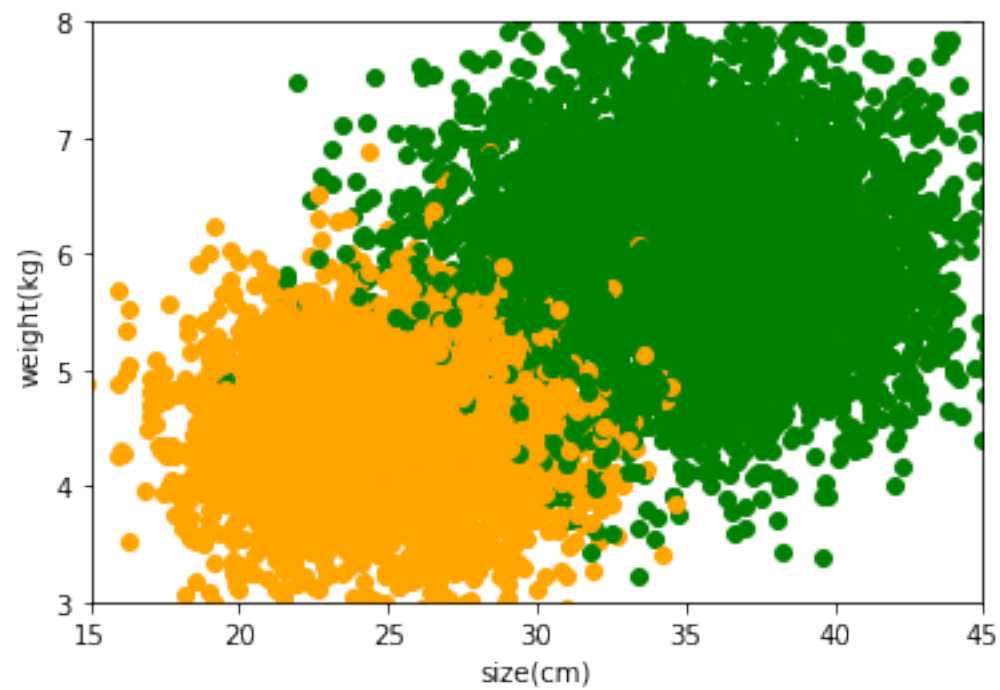
- 西瓜
- 哈密瓜



# 西瓜与哈密瓜分类器

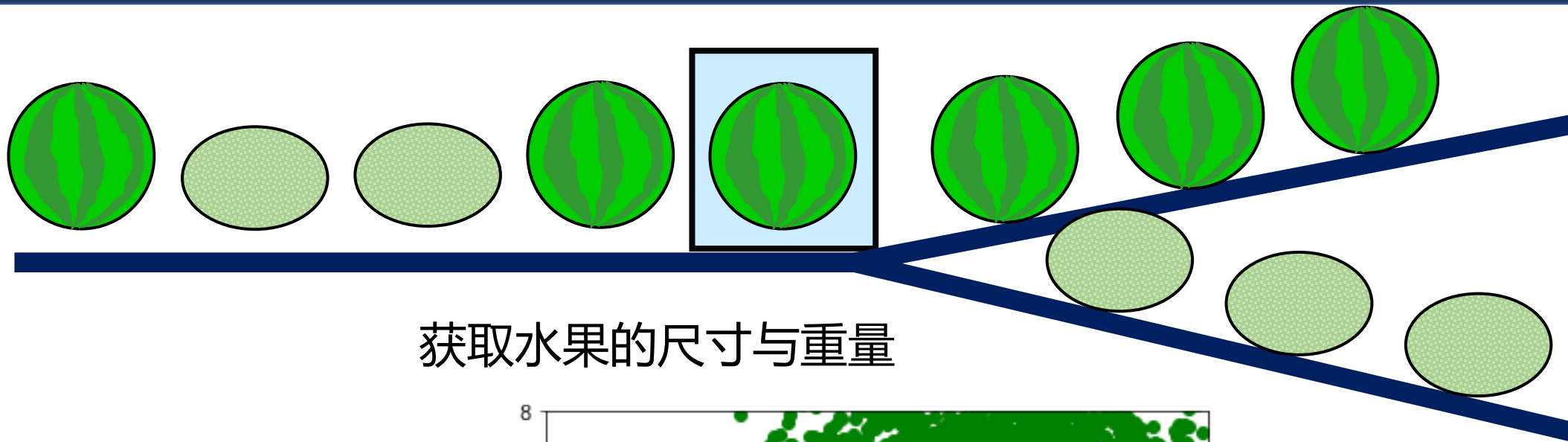


获取水果的尺寸与重量

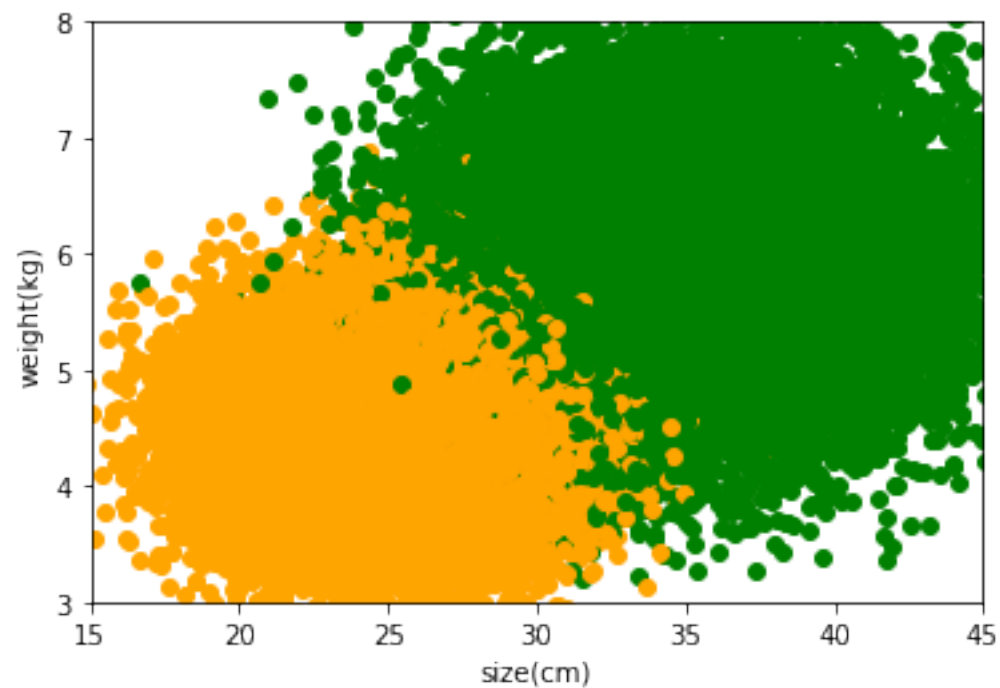


- 西瓜
- 哈密瓜

# 西瓜与哈密瓜分类器

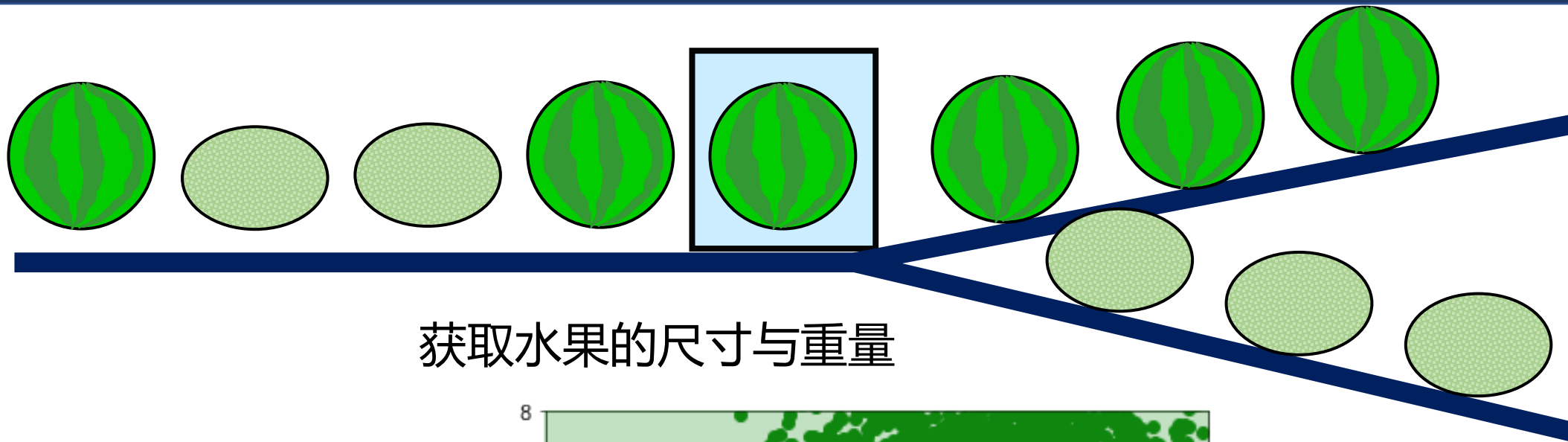


获取水果的尺寸与重量

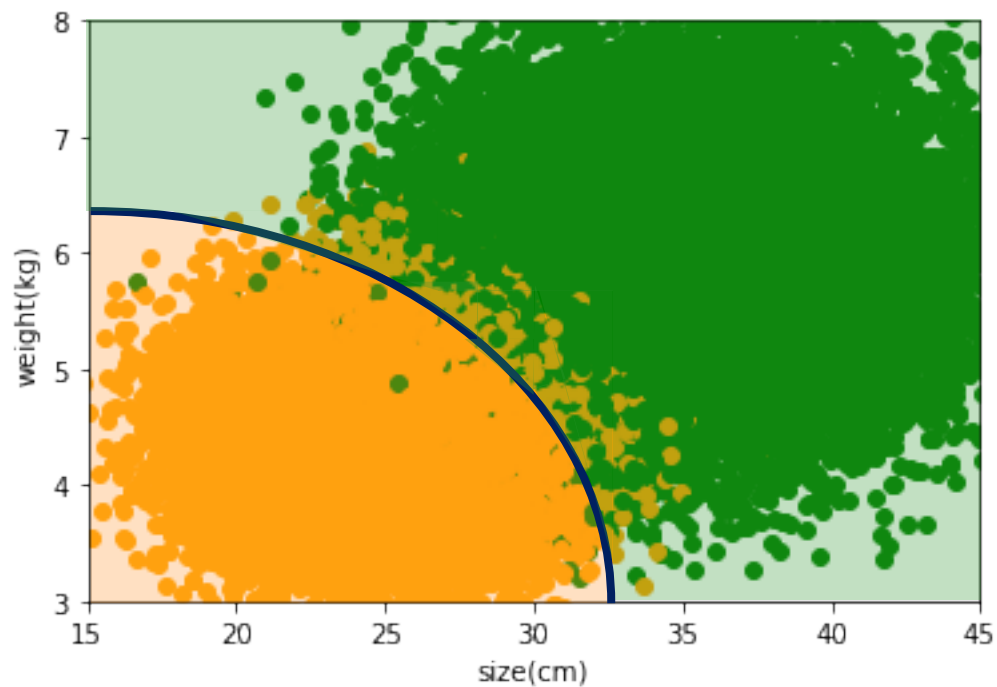


- 西瓜
- 哈密瓜

# 西瓜与哈密瓜分类器

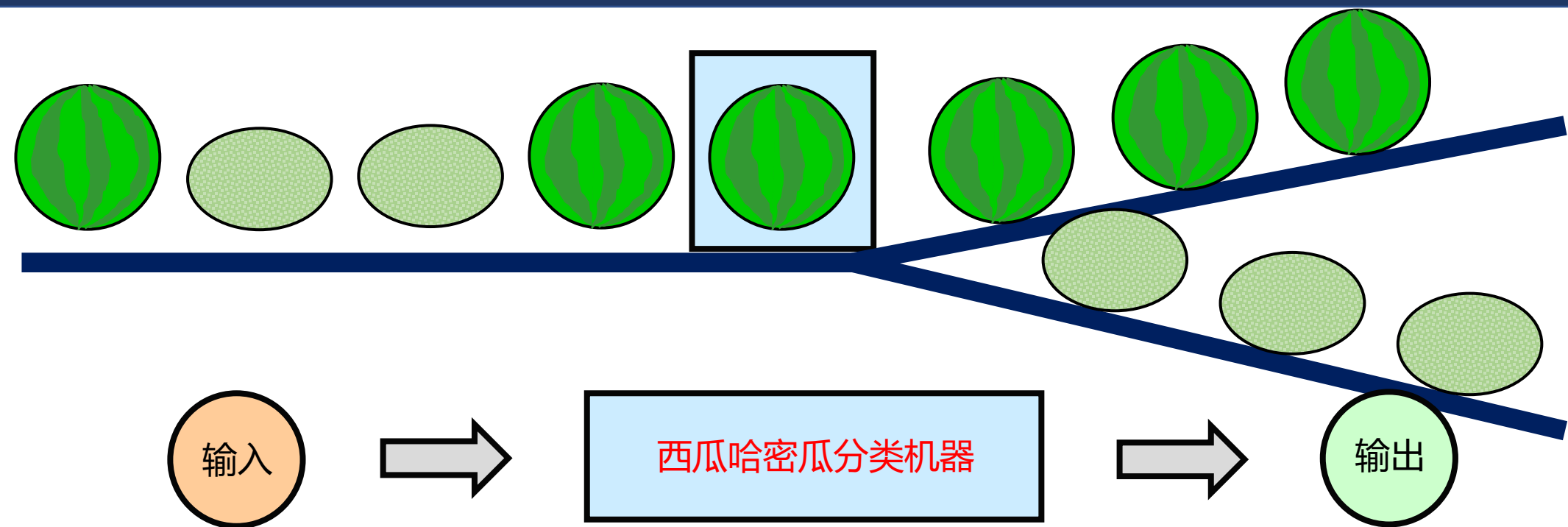


获取水果的尺寸与重量

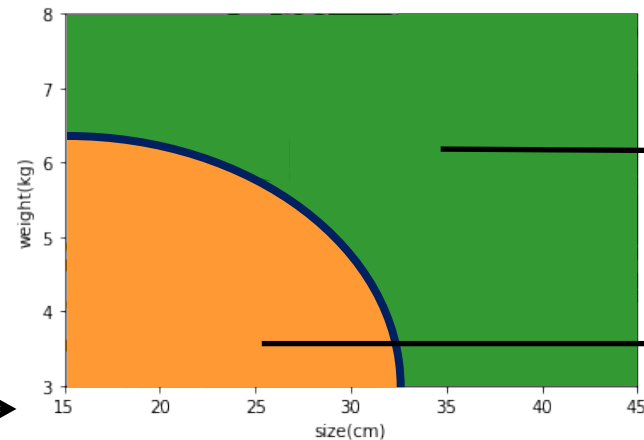


- 西瓜
- 哈密瓜

# 机器学习过程



重量  
尺寸



西瓜或哈密瓜

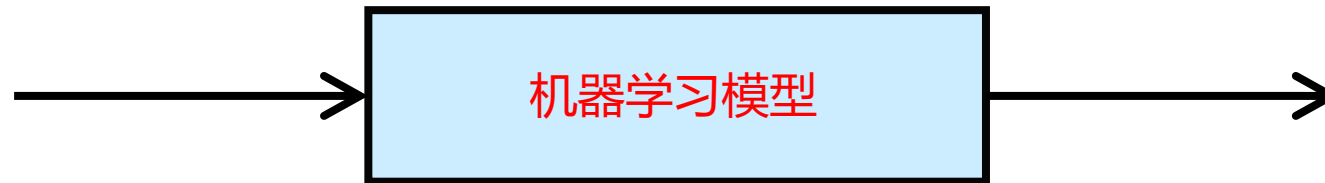
类标签

特征、描述符

# 一维模型

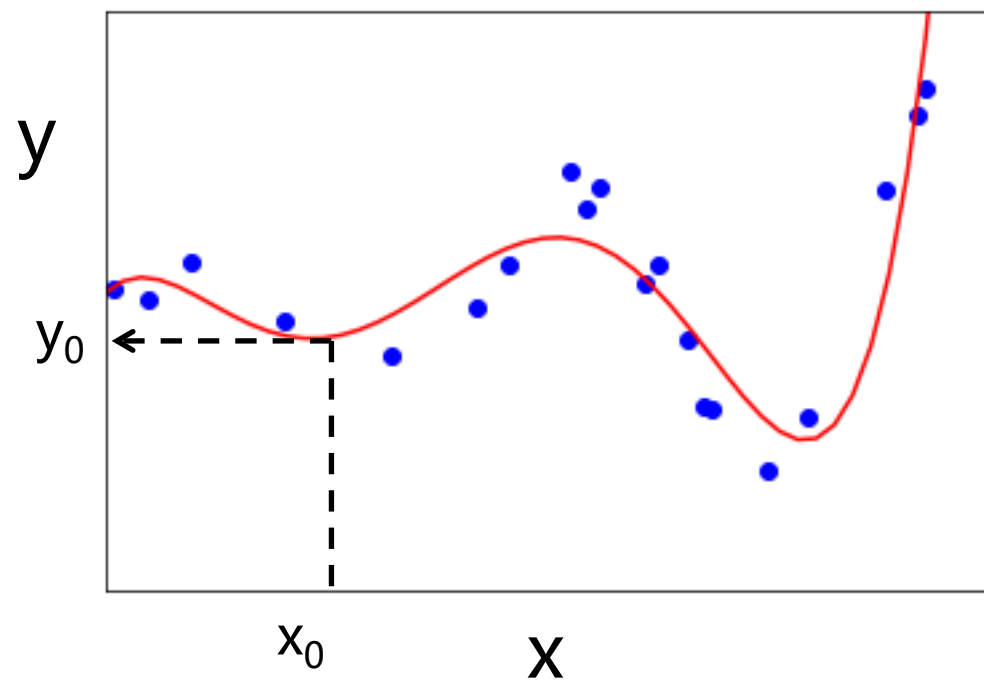
输入

$x$



输出

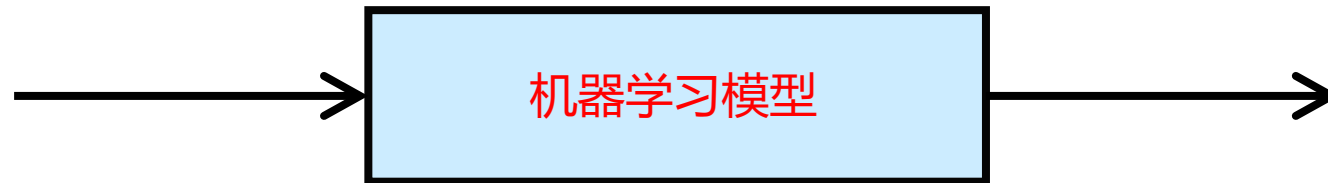
$y$



# 二维模型

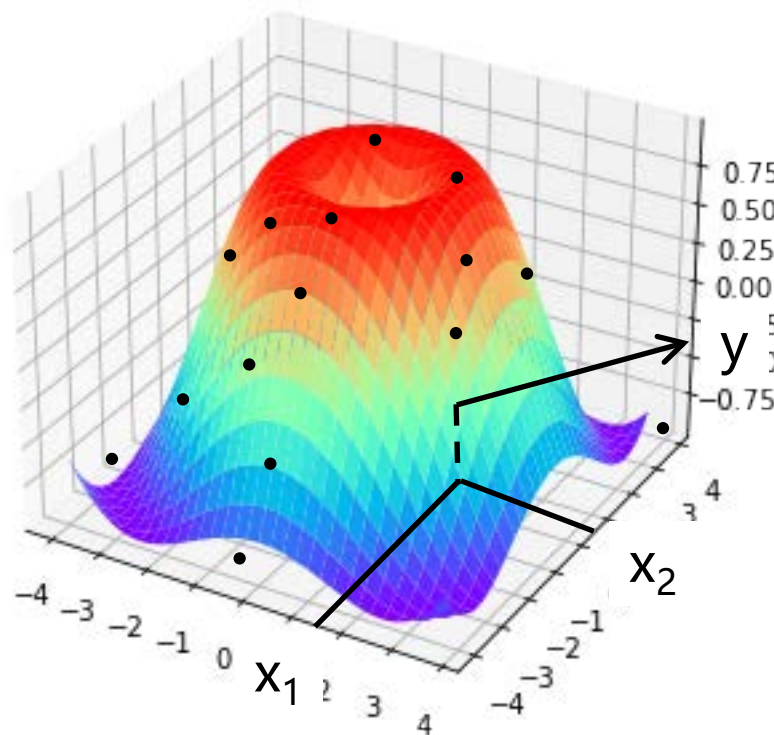
输入

$x_1, x_2$

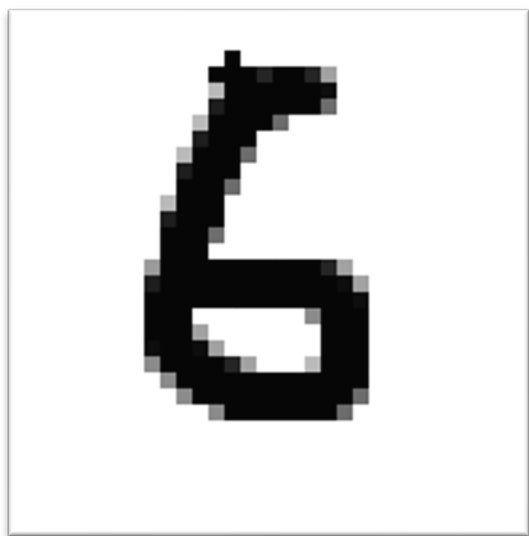


输出

$y$



# 高维模型



32×32

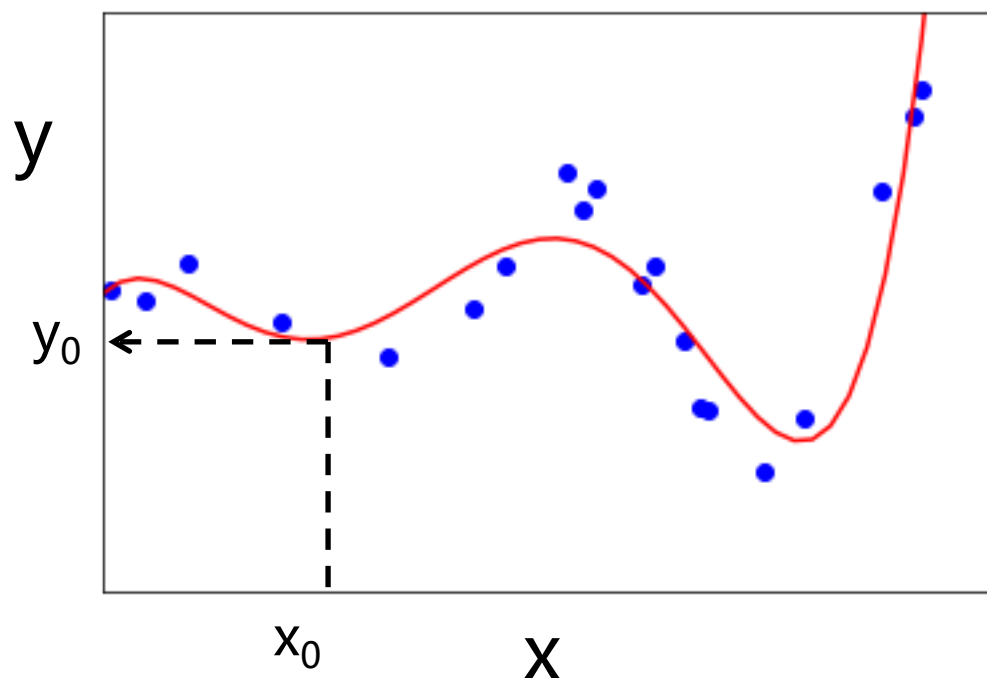


6

寻找高维到低维的映射关系

# 机器学习预测

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



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？



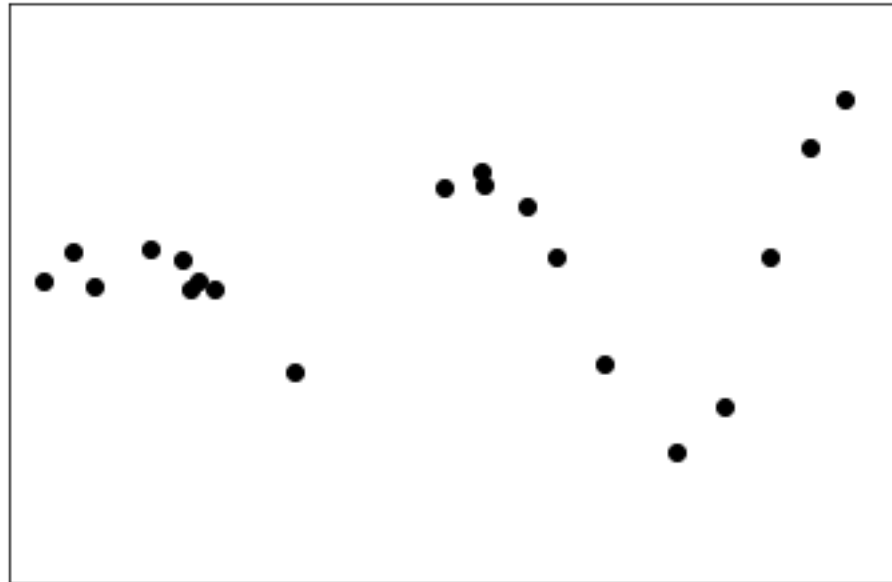
函数拟合插值



但是，高维函数的拟合并没有想象的简单

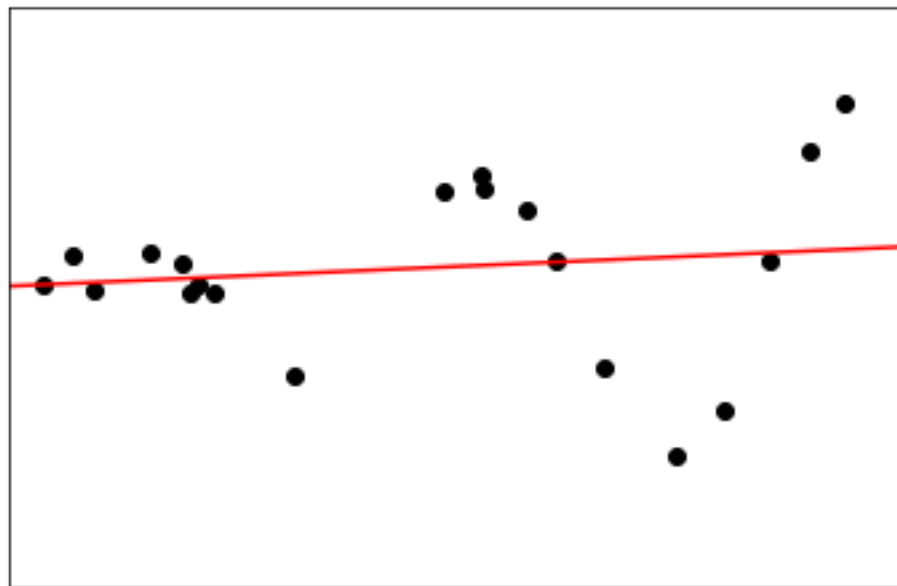
为什么机器学习比想象的复杂？

以曲线拟合为例



为什么机器学习比想象的复杂？

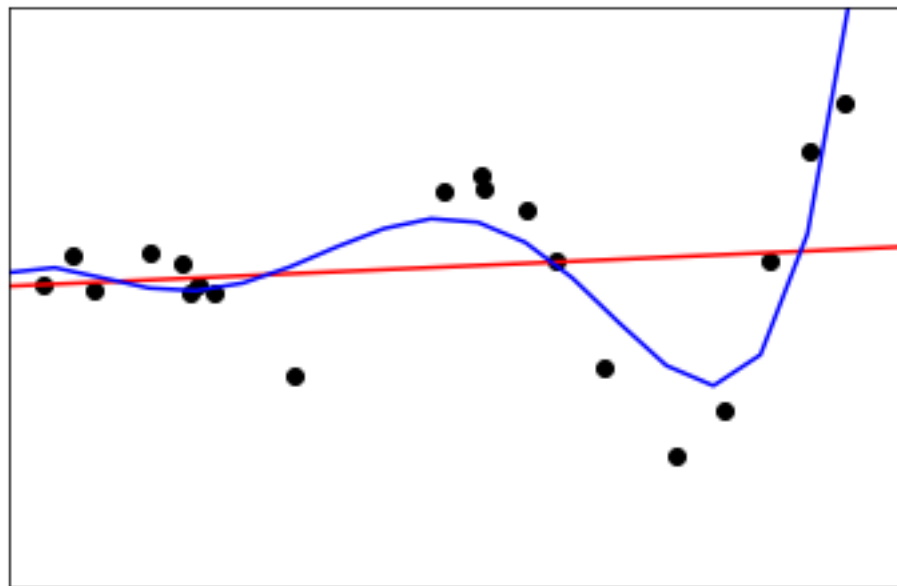
以曲线拟合为例



模型1

为什么机器学习比想象的复杂？

以曲线拟合为例

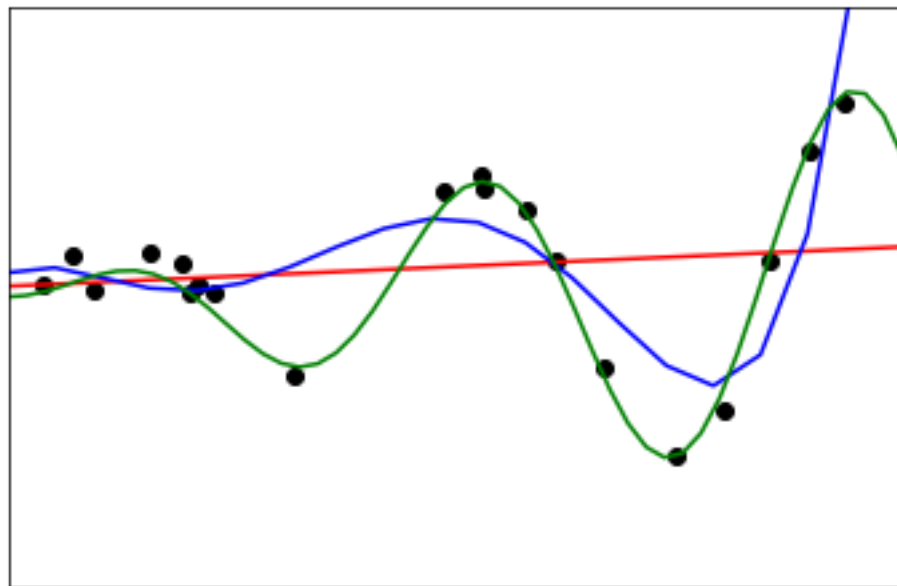


模型2

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

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

**nature**

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [articles](#) > article

Article | Published: 08 July 2020

## A mobile robotic chemist

[Benjamin Burger](#),  
[Wang, Xiaobo Li](#),  
[Sprick](#) & [Andrew](#)

*Nature* **583**, 237–241 (2020) | DOI: 10.1038/s41586-020-20342-6

47k Accesses | 1 Citation

### Abstract

Technologies such as machine learning and high-throughput experimentation that are defined by their length scale can

**nature COMMUNICATIONS**

ARTICLE

<https://doi.org/10.1038/s41467-020-20342-6> OPEN

## Machine learned features predict accurate adsorption

Victor Fung<sup>1</sup>, Guoxiang Hu<sup>2</sup>, P. Ganesh<sup>3</sup>

Materials databases generated by high-throughput density functional theory (DFT), have become valuable tools for predicting catalytic activity. However, heterogeneous catalysts, though the computational cost presents a crucial roadblock. Hence there is a significant need for machine learning models that can predict catalytic features, in lieu of DFT, to accurately predict catalytic activity. Here, we demonstrate an approach to predict energy barriers for catalytic reactions using a machine learning model to automatically extract features from density of states (DOS). The model, DOSnet, is evaluated on a set of surfaces, yielding a mean absolute error on the order of 0.1 eV. This model provides physically meaningful predictions and insights into the electronic structure without additional computational cost for the accelerated discovery of materials and catalysts.

## REVIEW

<https://doi.org/10.1038/s41586-018-0337-2>

## Machine learning for molecular and materials science

Keith T. Butler<sup>1</sup>, Daniel W. Davies<sup>2</sup>, Hugh Cartwright<sup>3</sup>, Olexandr Isayev<sup>4\*</sup> & Aron Walsh<sup>5,6\*</sup>

Here we summarize techniques that are accelerating a future of machine learning accelerated by artificial intelligence.

The Schrödinger equation is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business

Machine learning (ML) (1) is the study of computer algorithms capable of learning to improve their performance of a task on the basis of their own previous experience. The field is closely related to pattern recognition and statistical inference. As an engineering field, ML has become steadily more mathematical and more successful in applications over the past 20 years. Learning approaches such as data clustering, neural network classifiers, and nonlinear regression have found surprisingly wide application in the practice of engineering, business

correlate surprisingly well with subsequent gene expression analysis (3). Postgenomic biology prominently features large-scale gene expression data analyzed by clustering methods (4), a standard topic in unsupervised learning. Many other examples can be given of learning and pattern recognition applications in science. Where will this trend lead? We believe it will lead to appropriate, partial automation of every element of scientific method, from hypothesis generation to model construction to decisive experimentation. Thus, ML has the potential to

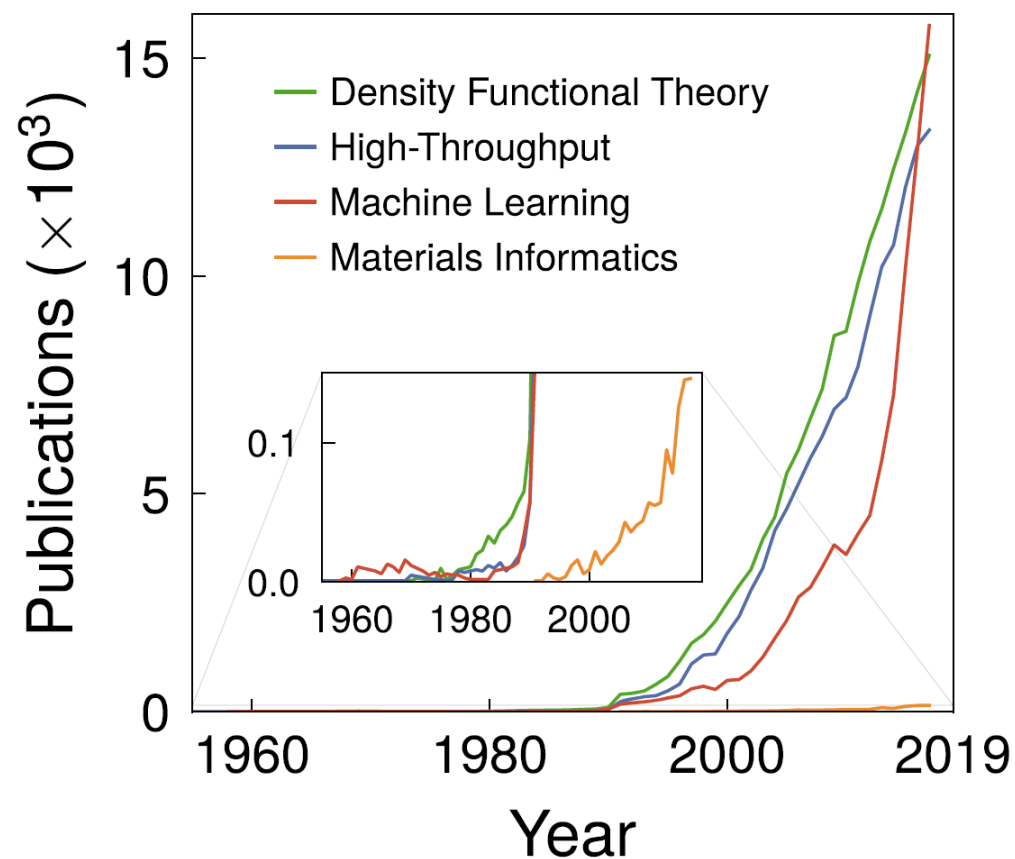
## Machine Learning for Science: State of the Art and Future Prospects

Eric Mjølness\* and Dennis DeCoste

Recent advances in machine learning methods, along with successful applications across a wide variety of fields such as planetary science and bioinformatics, promise powerful new tools for practicing scientists. This viewpoint highlights some useful characteristics of modern machine learning methods and their relevance to scientific applications. We conclude with some speculations on near-term progress and promising directions.

creating hypotheses, testing by decisive experiment or observation, and iteratively building up comprehensive testable models or theories is shared across disciplines. For each stage of this abstracted scientific process, there are relevant developments in ML, statistical inference, and pattern recognition that will lead to semiautomatic support tools of unknown but potentially broad applicability.

Increasingly, the early elements of scientific method—observation and hypothesis generation—face high data volumes, high data acquisition rates, or requirements for objective analysis that cannot be handled by human perception alone. This has been the situation in experimental particle physics for decades. There automatic pattern recognition for significant events is well developed, including Hough transforms, which are foundational in pattern



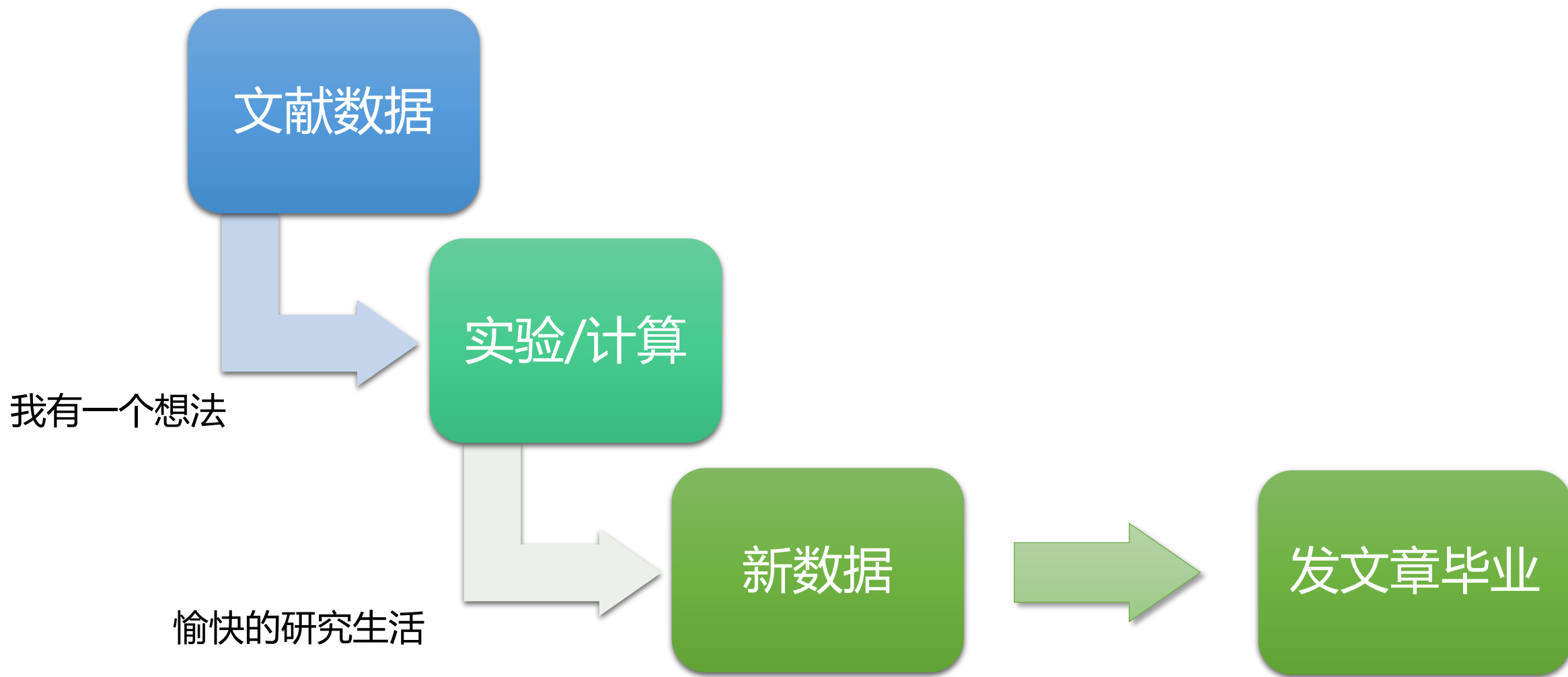
斯坦福大学2021年AI数据生成索引报告:

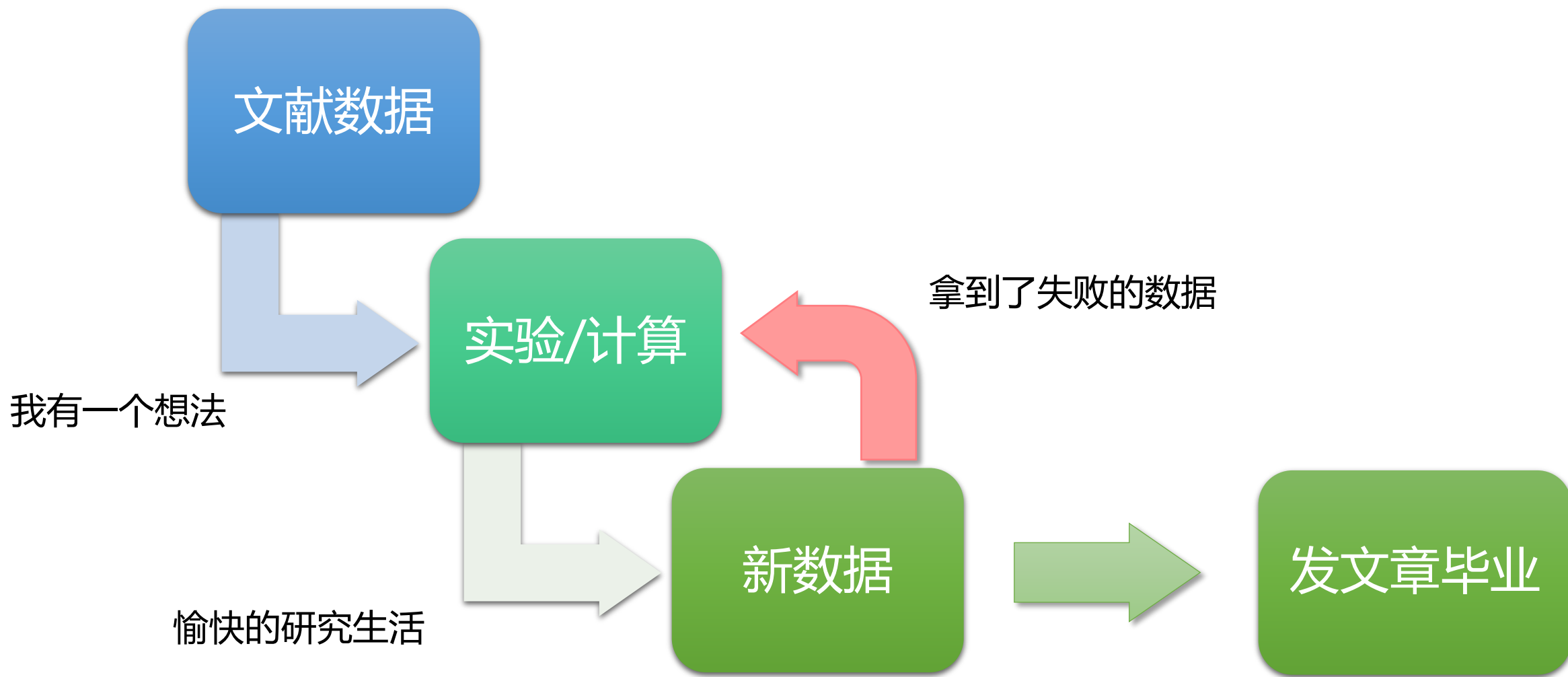
涉及AI的期刊出版物数量于2019年到2020年之间增长了**34.5%**，而在2018年到2019年此增长率仅为**19.6%**。

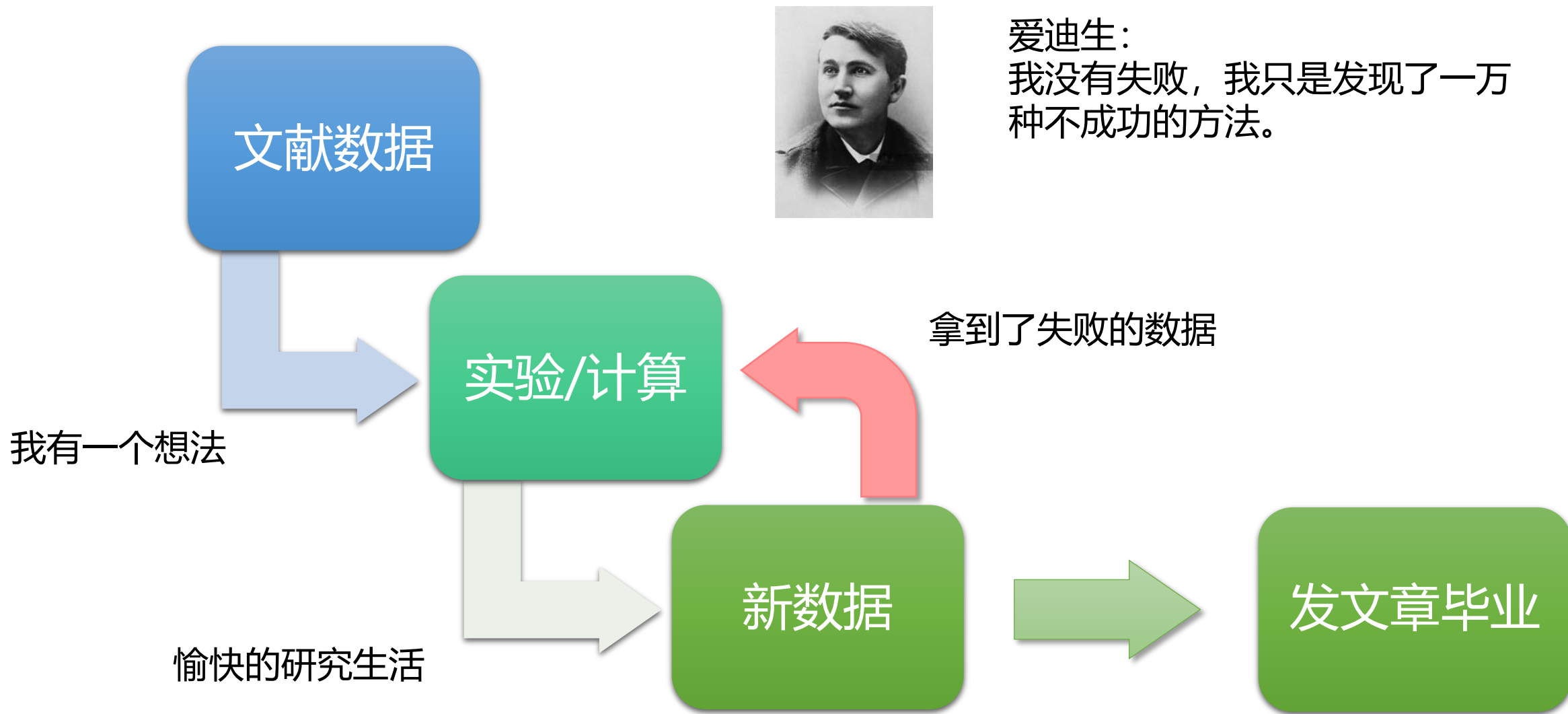
2019年，人工智能类出版物占全球所有经同行评审科学类出版物的**3.8%**，相较2011年的**1.3%**有所提升。

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



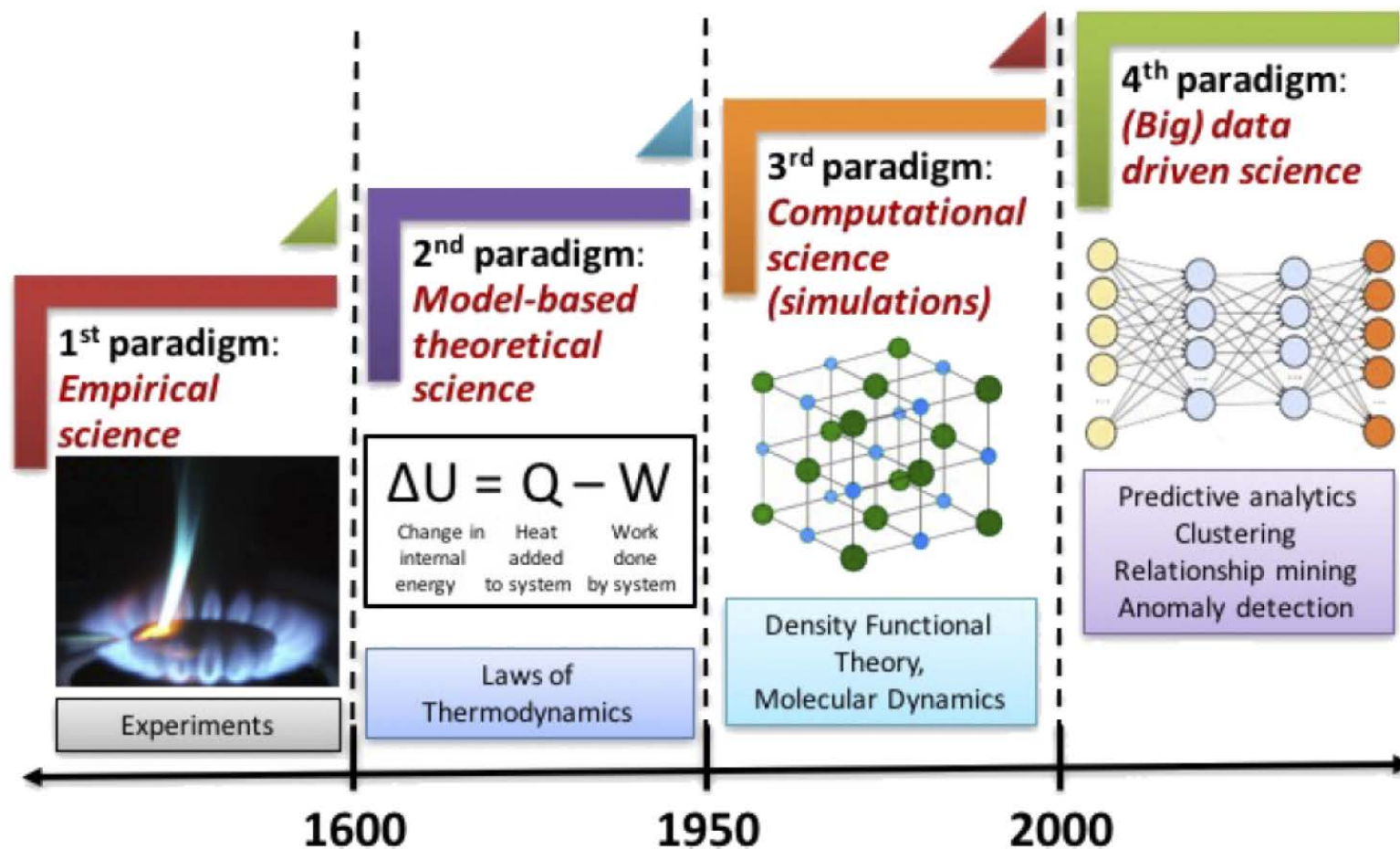


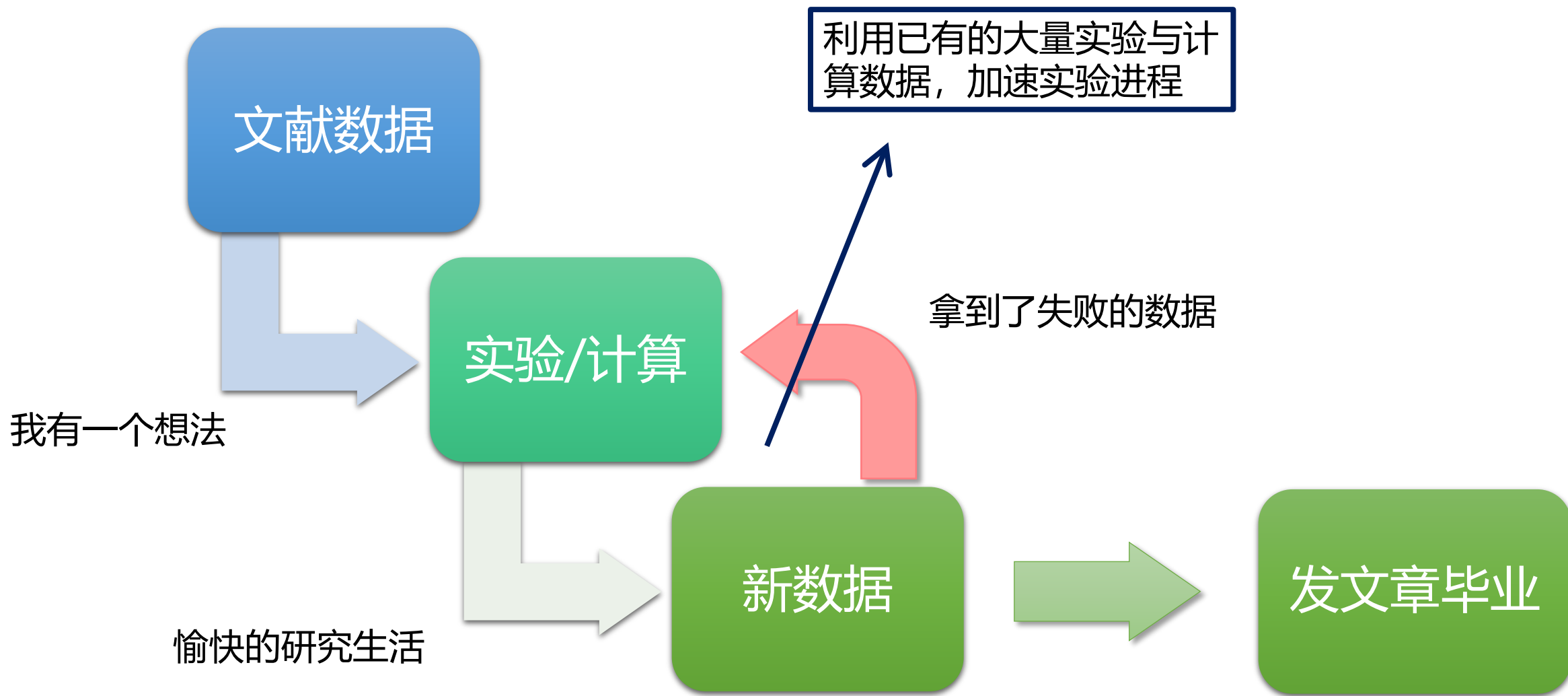








# 时代变了

但是，现在已经是2022年了。。。





1. 机器学习是什么
2. 机器学习与化学研究
3. 机器学习库

				
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、易用、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

<https://scikit-learn.org>

## scikit-learn

*Machine Learning in Python*

Getting Started

Release Highlights for 1.0

GitHub

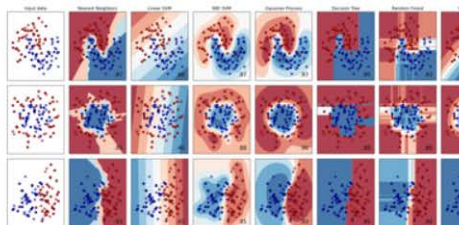
- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

### Classification

Identifying which category an object belongs to.

**Applications:** Spam detection, image recognition.

**Algorithms:** SVM, nearest neighbors, random forest, and more...



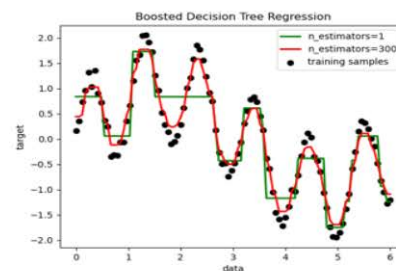
Examples

### Regression

Predicting a continuous-valued attribute associated with an object.

**Applications:** Drug response, Stock prices.

**Algorithms:** SVR, nearest neighbors, random forest, and more...



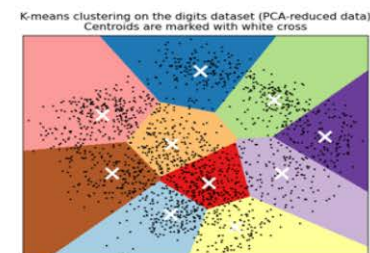
Examples

### Clustering

Automatic grouping of similar objects into sets.

**Applications:** Customer segmentation, Grouping experiment outcomes

**Algorithms:** k-Means, spectral clustering, mean-shift, and more...



Examples

### Dimensionality reduction

Reducing the number of random variables to consider.

**Applications:** Visualization, Increased efficiency

### Model selection

Comparing, validating and choosing parameters and models.

**Applications:** Improved accuracy via parameter tun-

### Preprocessing

Feature extraction and normalization.

**Applications:** Transforming input data such as text for use with machine learning algorithms.