Python与机器学习

——机器学习简介

华算科技 黄老师

2022年1月18日

华算科技
HUASUAN TECHNOLOGY

调查

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

调查

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

目录

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

机器学习

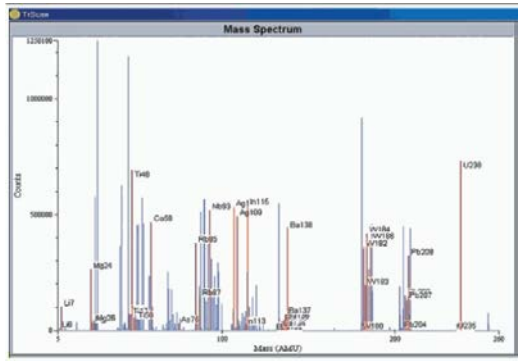
电脑程序

通过经验自动改进

Machine Learning

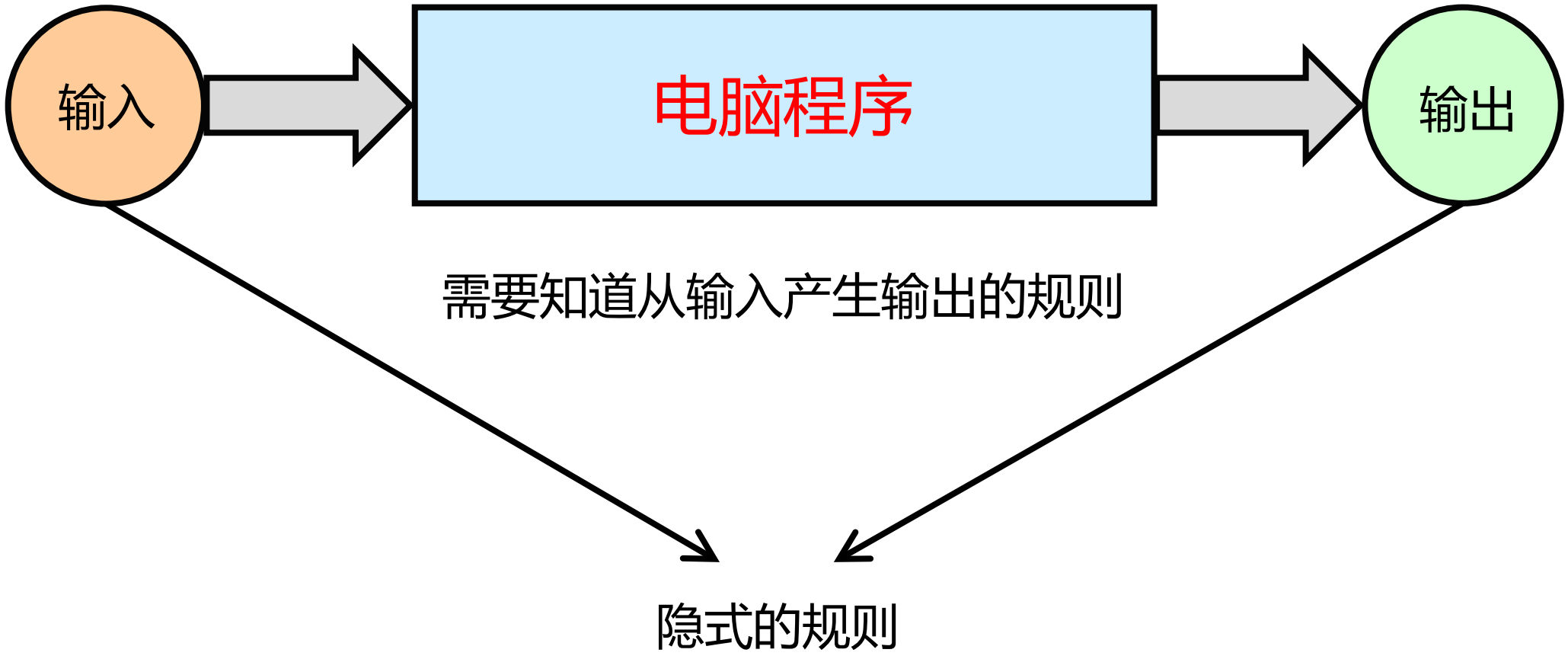
电脑程序?

电脑程序即为处理输入获得输出的方式

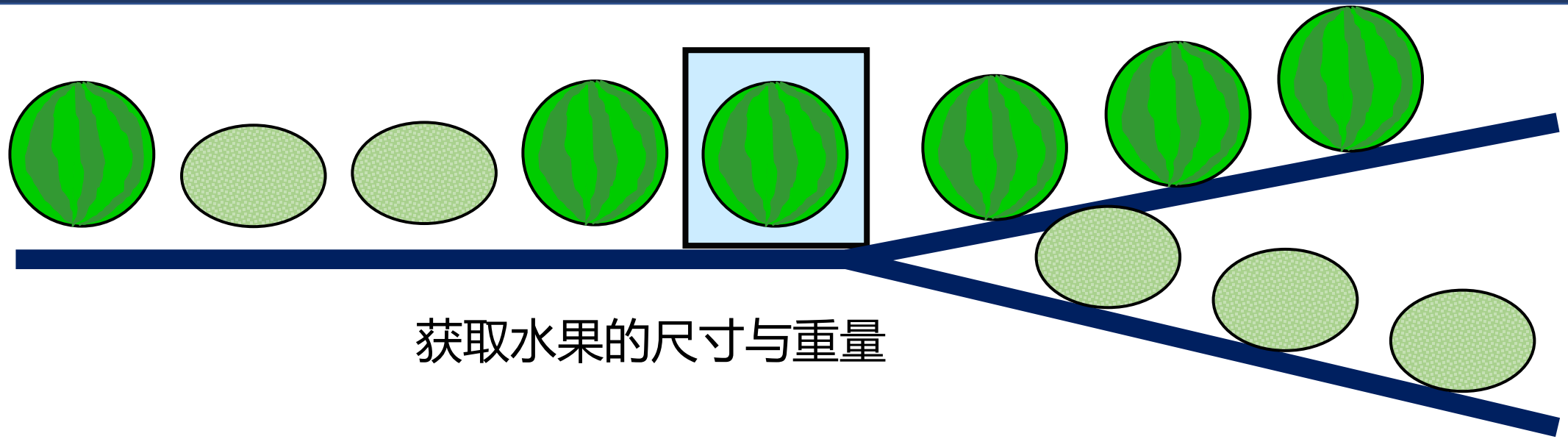


电脑程序?

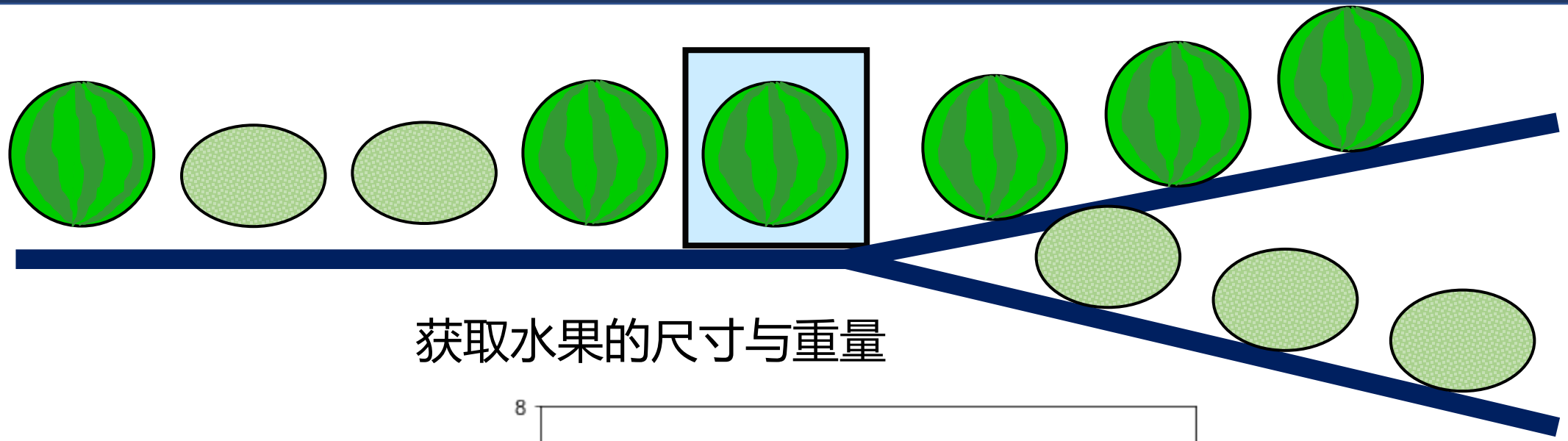
电脑程序即为处理输入获得输出的方式



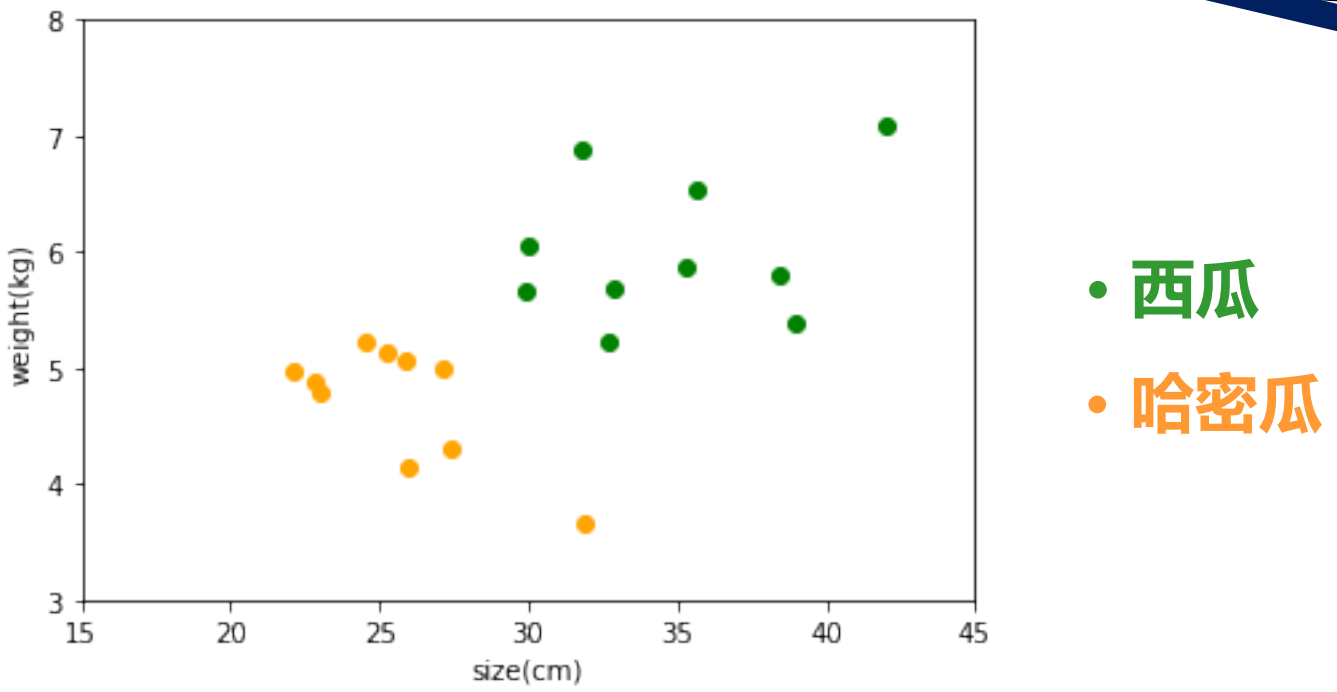
西瓜与哈密瓜分类器



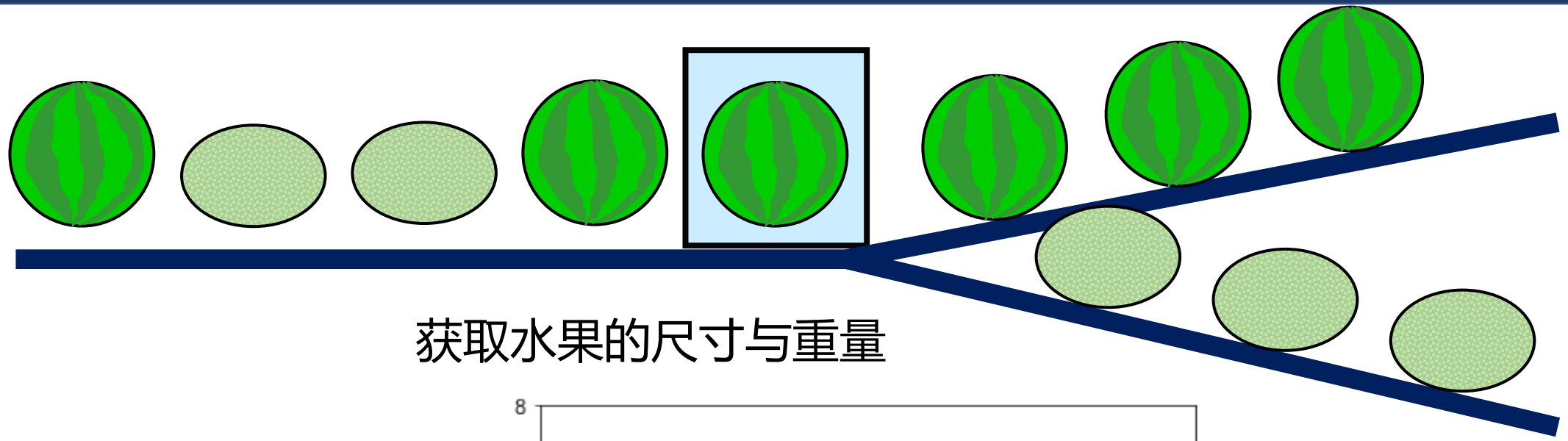
西瓜与哈密瓜分类器



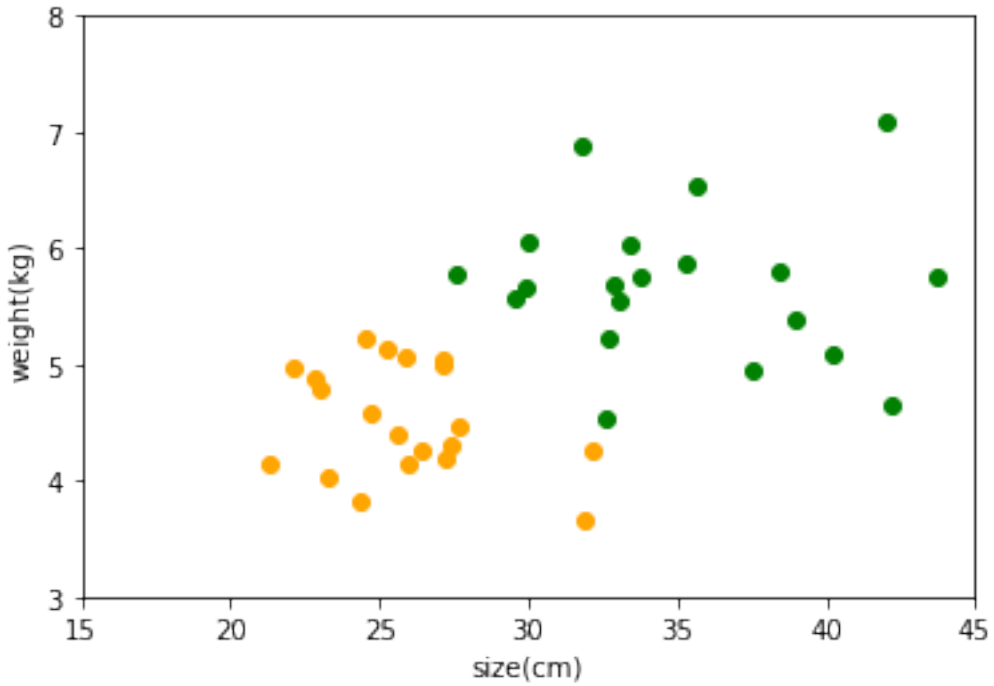
获取水果的尺寸与重量



西瓜与哈密瓜分类器

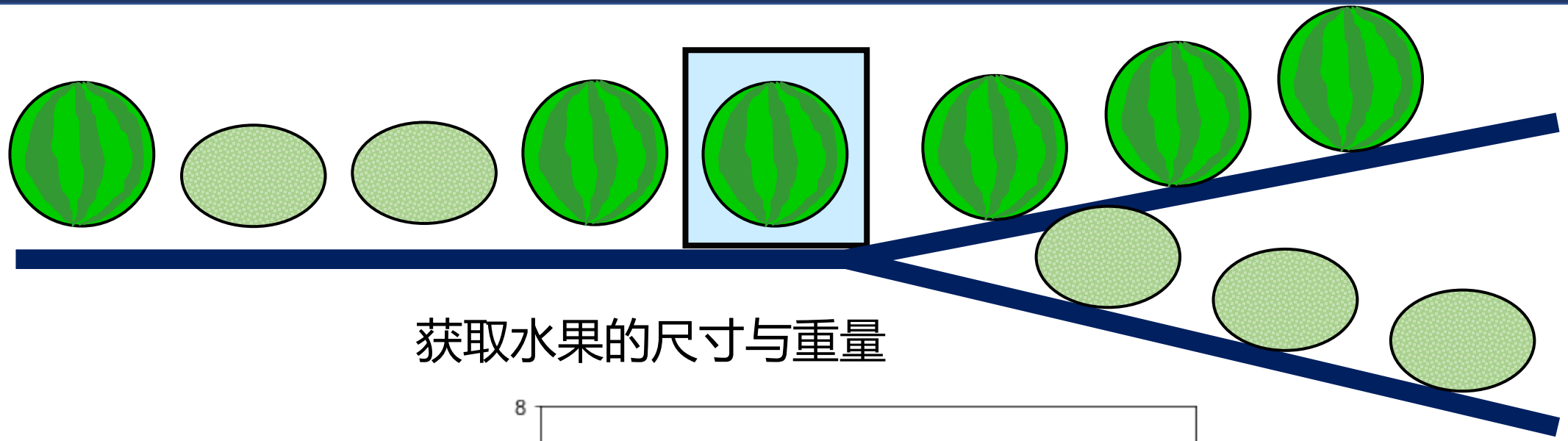


获取水果的尺寸与重量

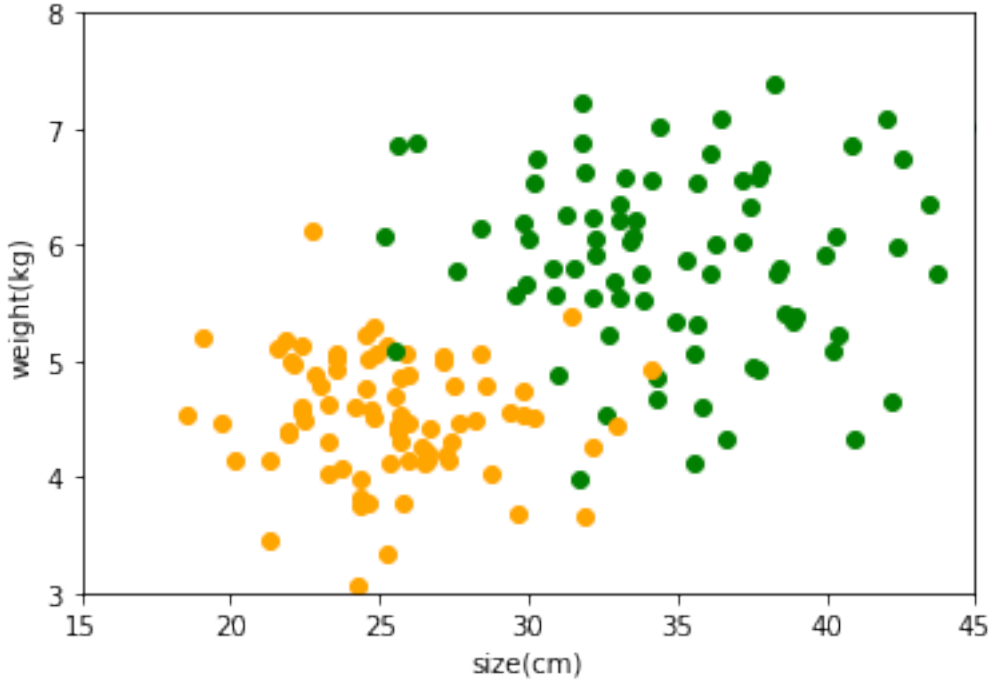


- 西瓜
- 哈密瓜

西瓜与哈密瓜分类器

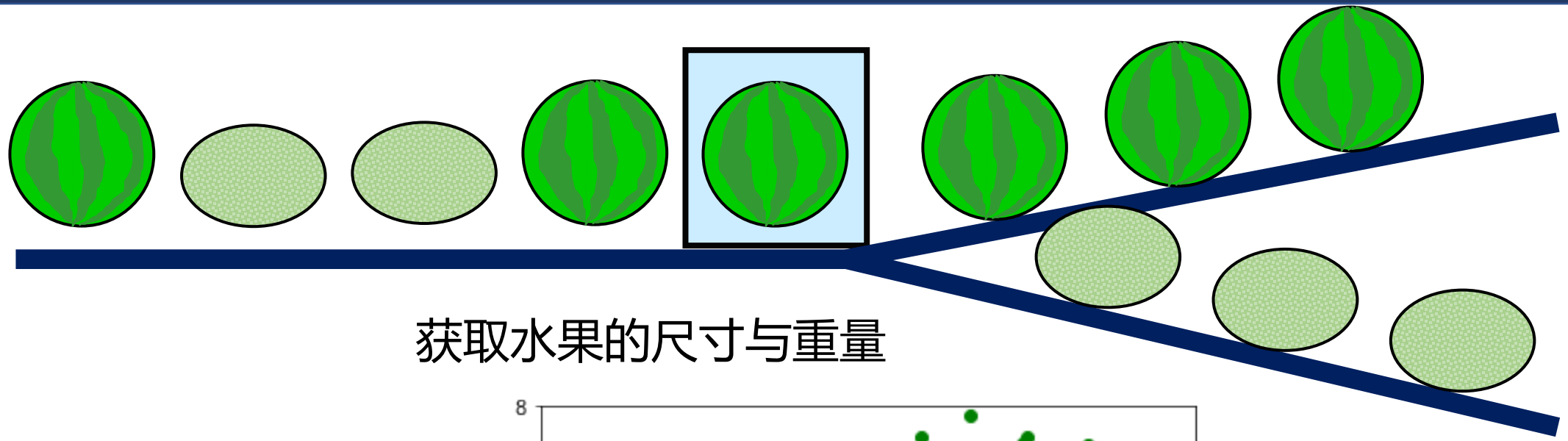


获取水果的尺寸与重量

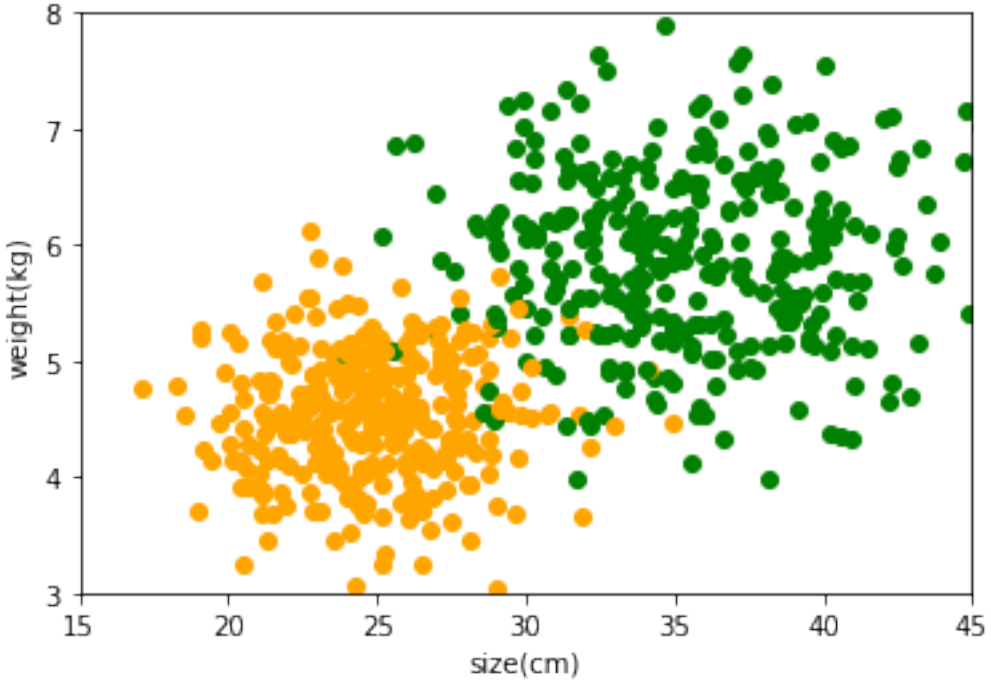


- 西瓜
- 哈密瓜

西瓜与哈密瓜分类器

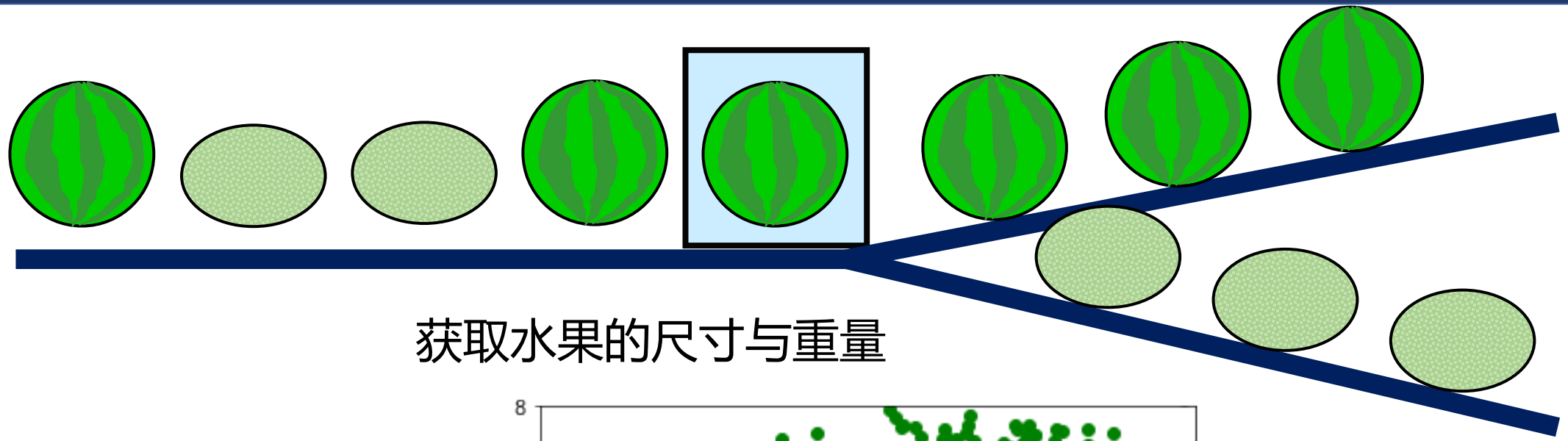


获取水果的尺寸与重量

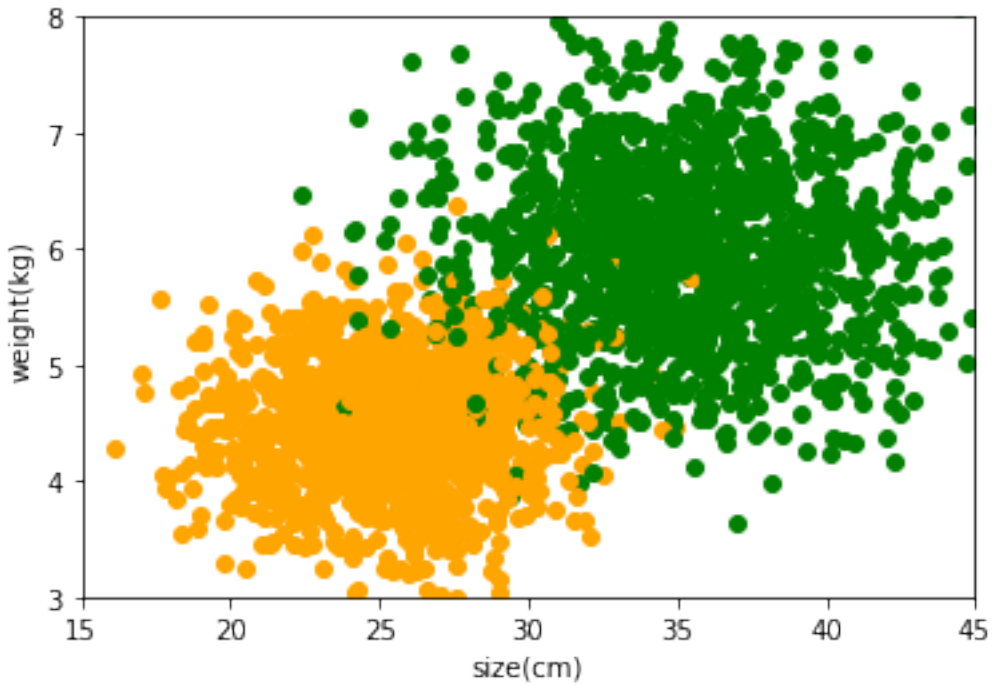


- 西瓜
- 哈密瓜

西瓜与哈密瓜分类器

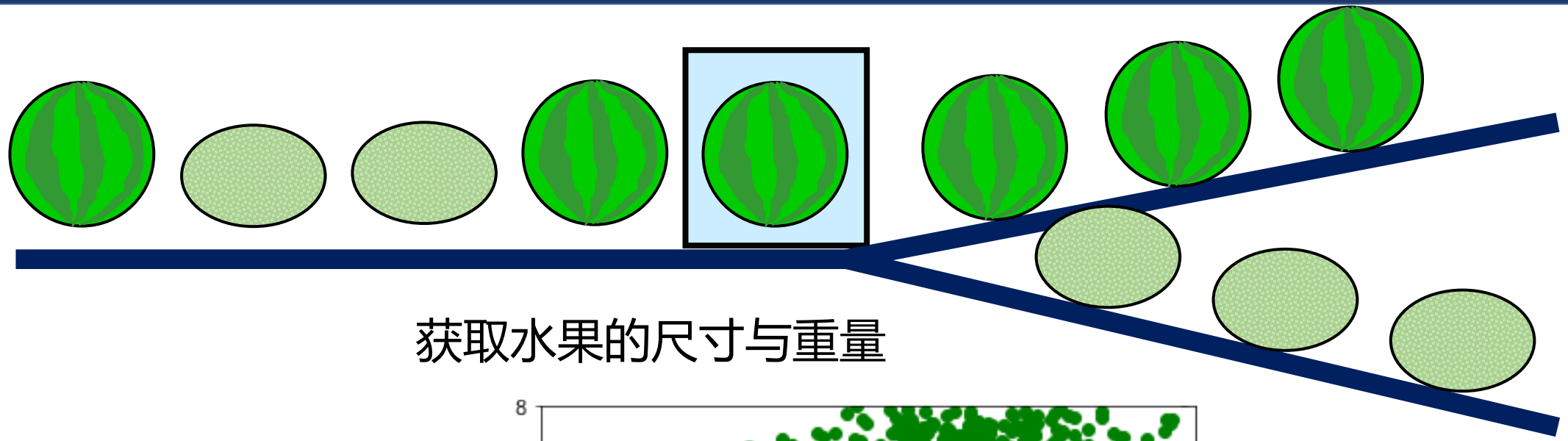


获取水果的尺寸与重量

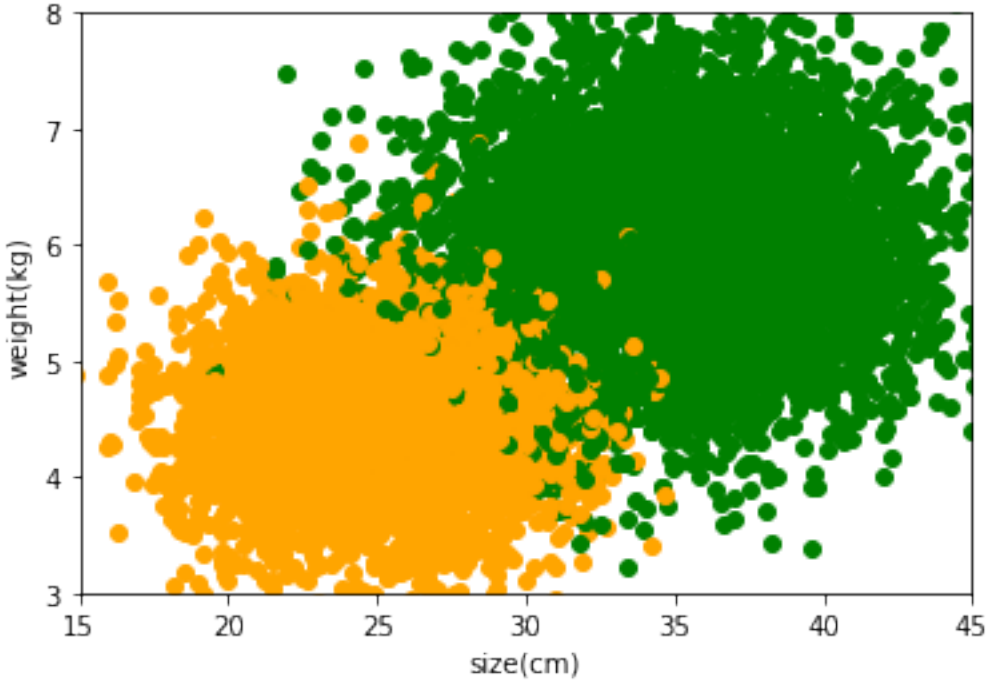


- 西瓜
- 哈密瓜

西瓜与哈密瓜分类器

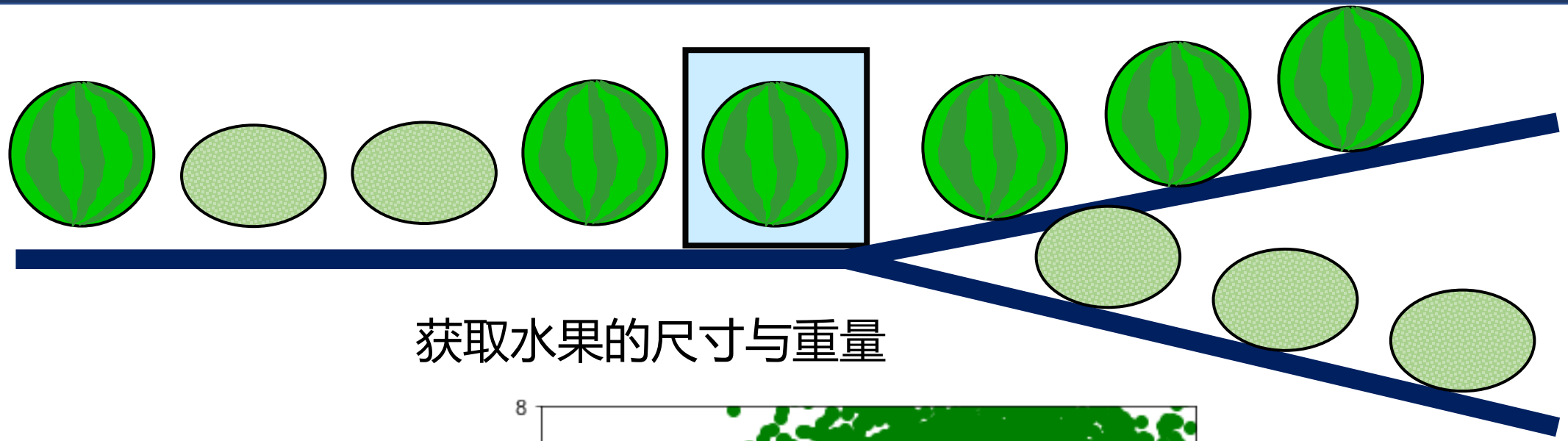


获取水果的尺寸与重量

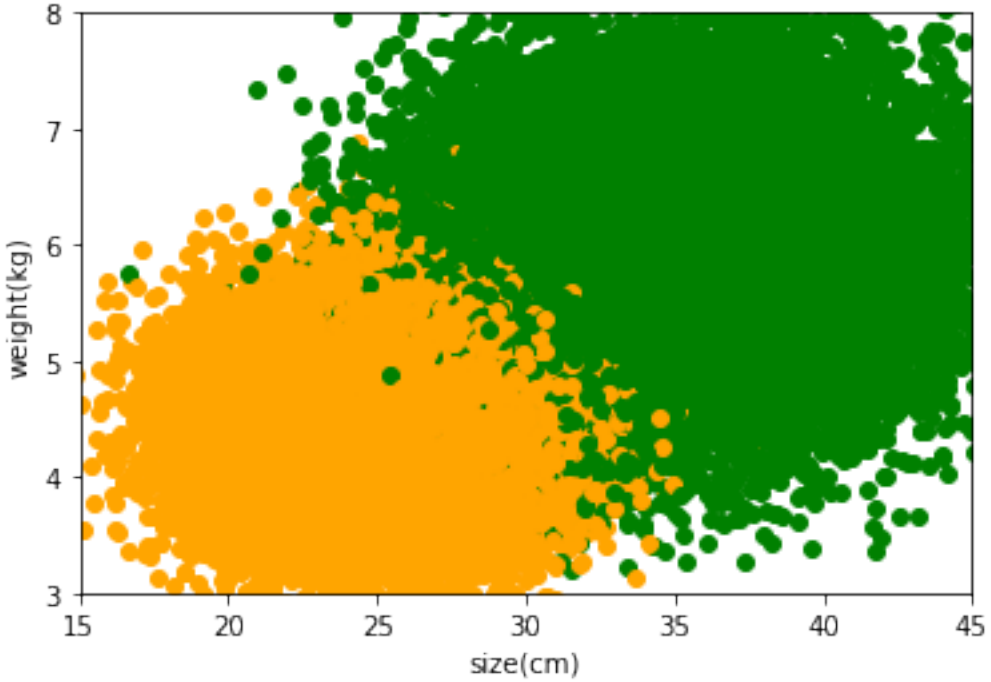


- 西瓜
- 哈密瓜

西瓜与哈密瓜分类器

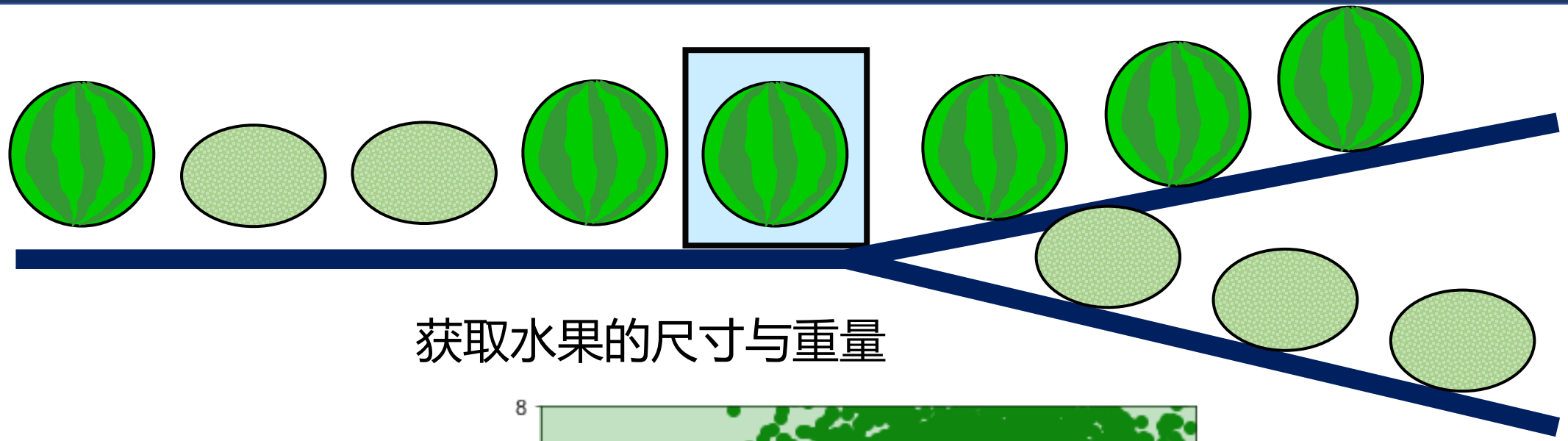


获取水果的尺寸与重量

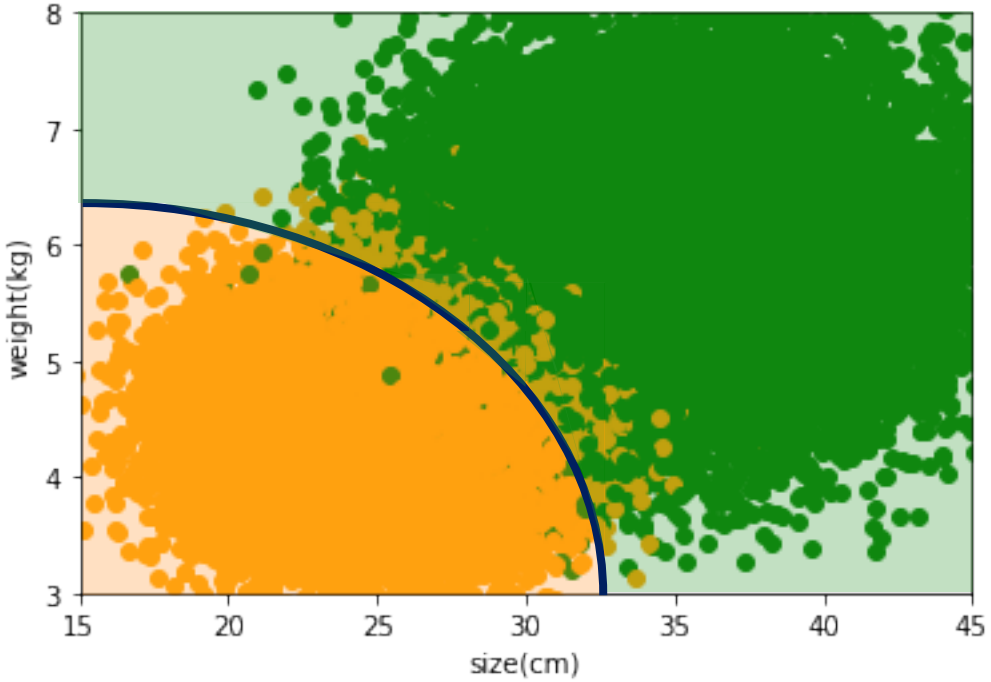


- 西瓜
- 哈密瓜

西瓜与哈密瓜分类器

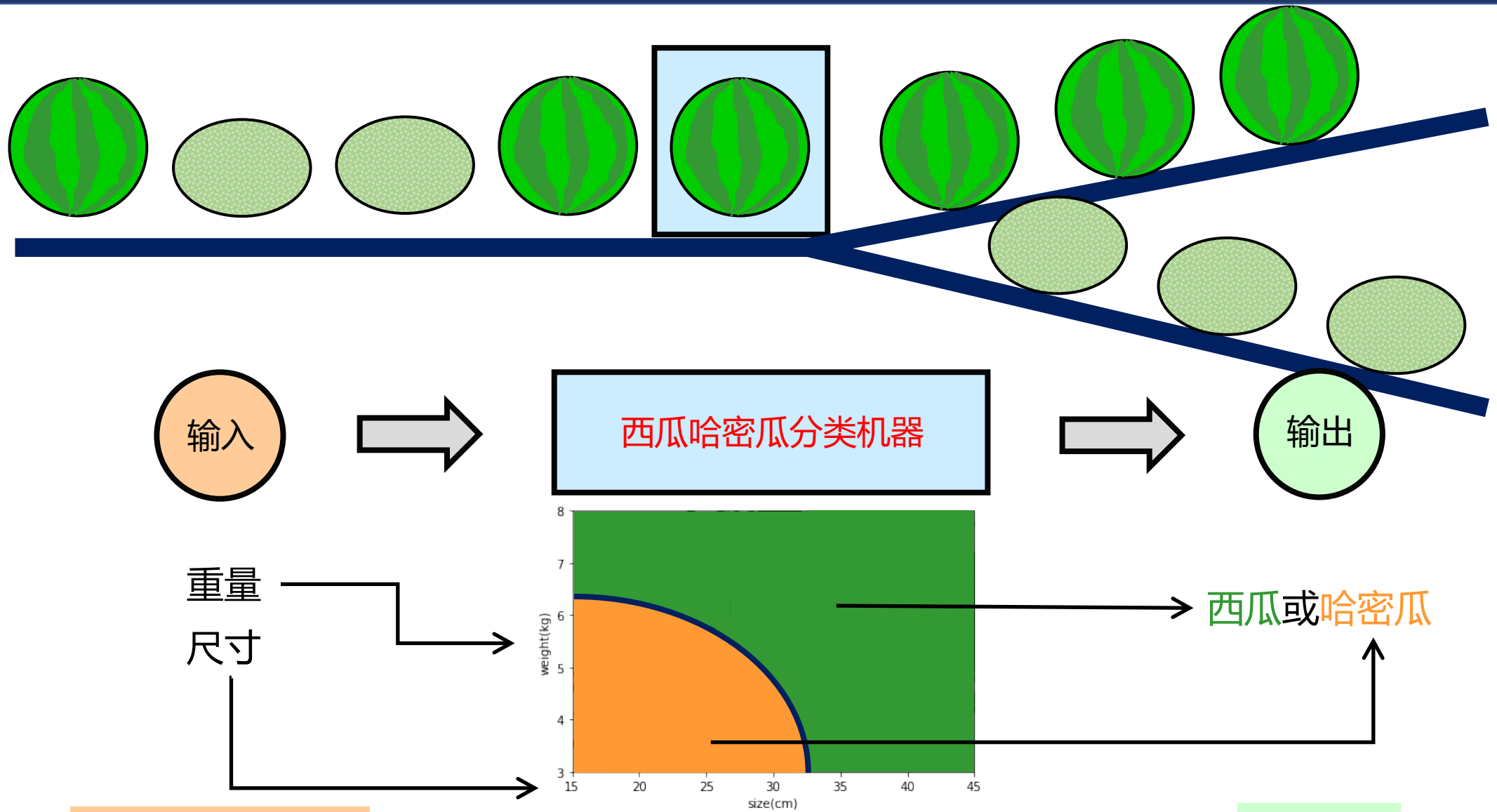


获取水果的尺寸与重量

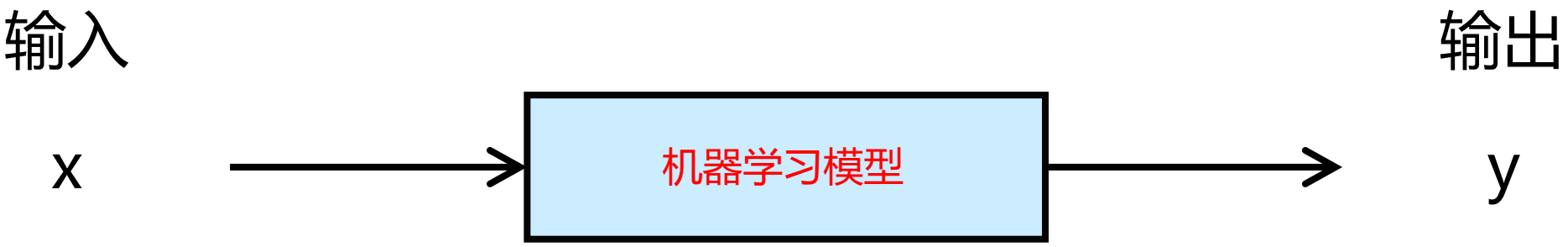


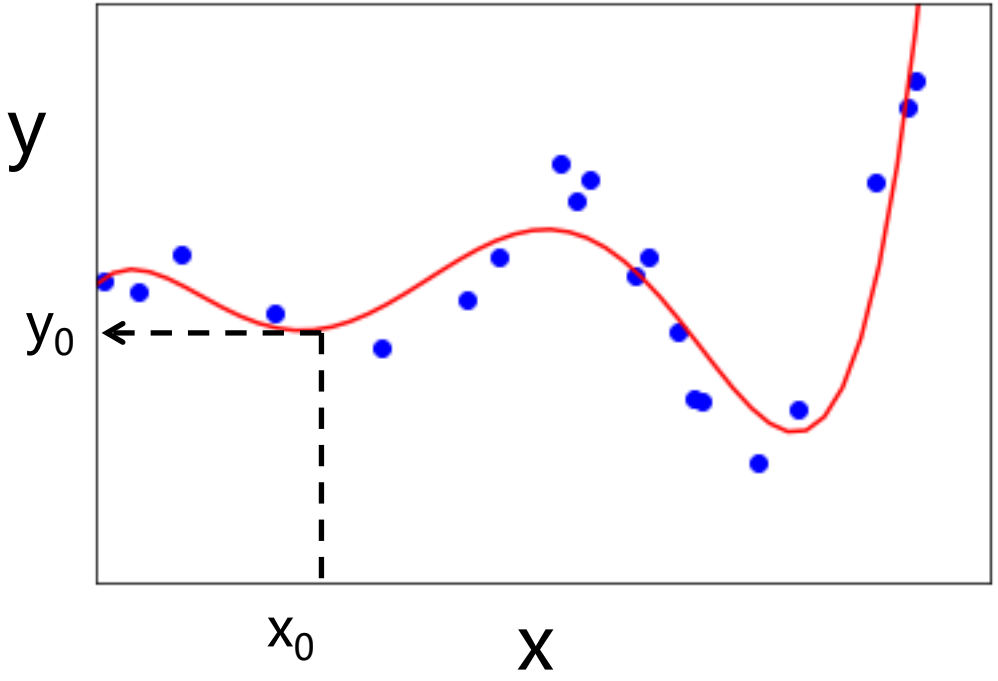
- 西瓜
- 哈密瓜

机器学习过程



一维模型





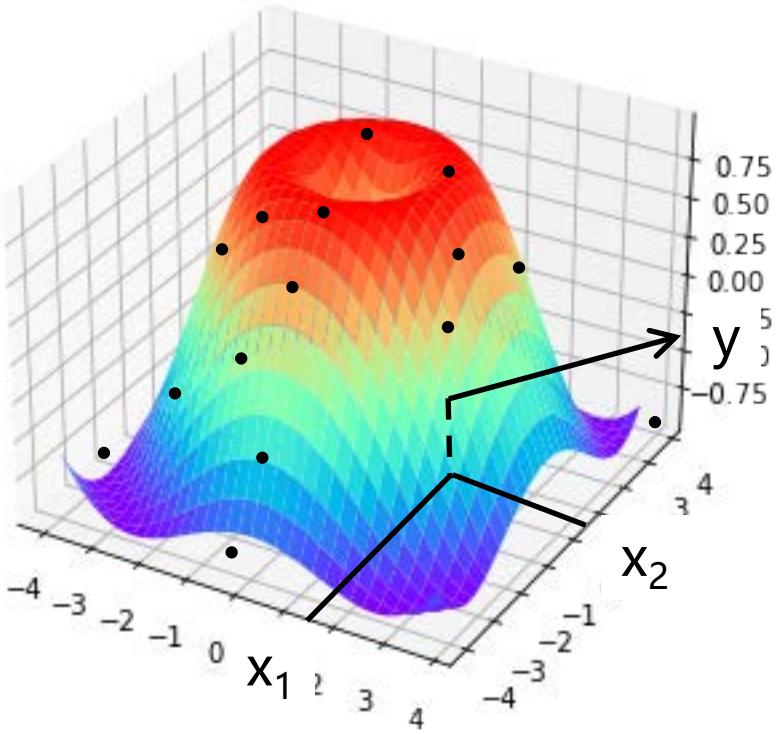
二维模型

输入

x_1, x_2

输出

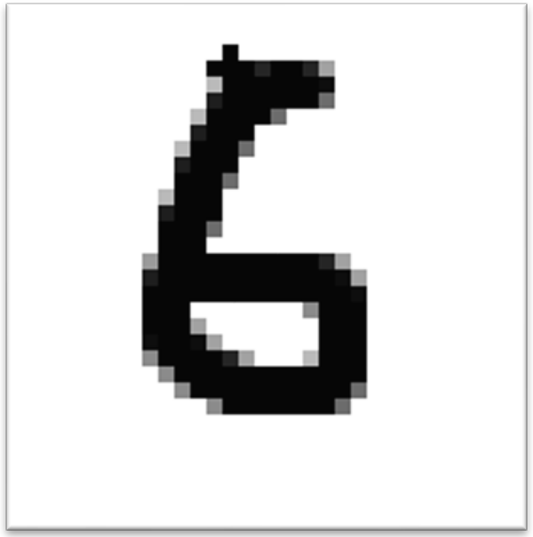
y



高维模型

输入 → 机器学习模型 → 输出

$x_1, x_2, \ldots, x_n$ → 机器学习模型 → y



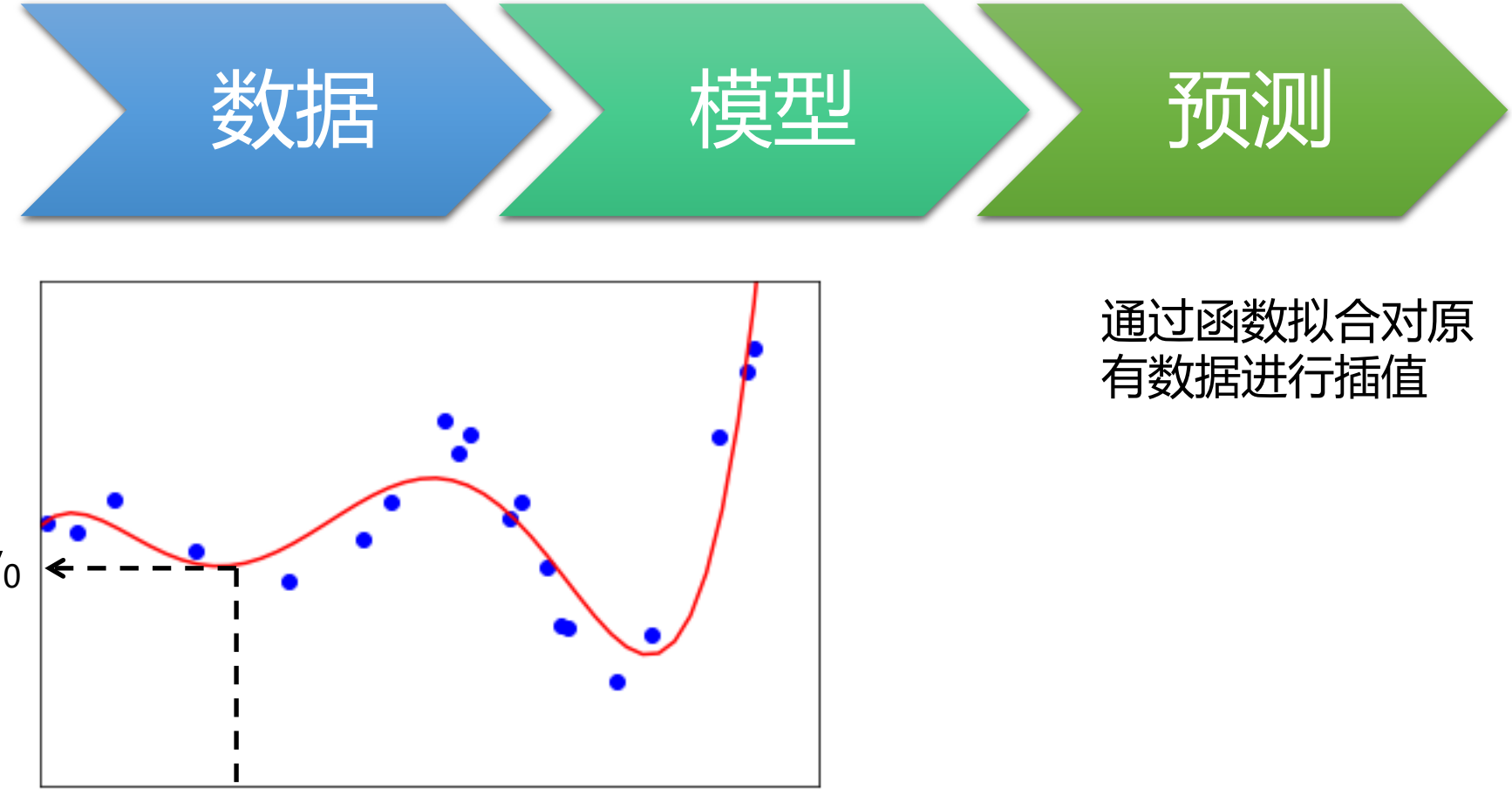
32×32

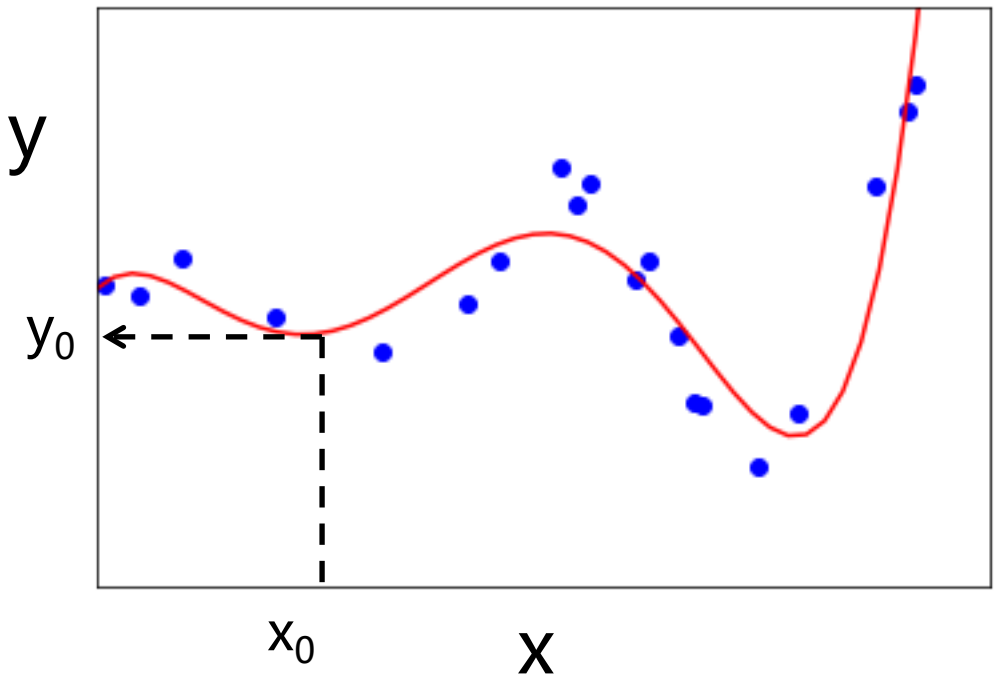
→ 6

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

机器学习本质

魔法?

机器像人一样的思考? ✗

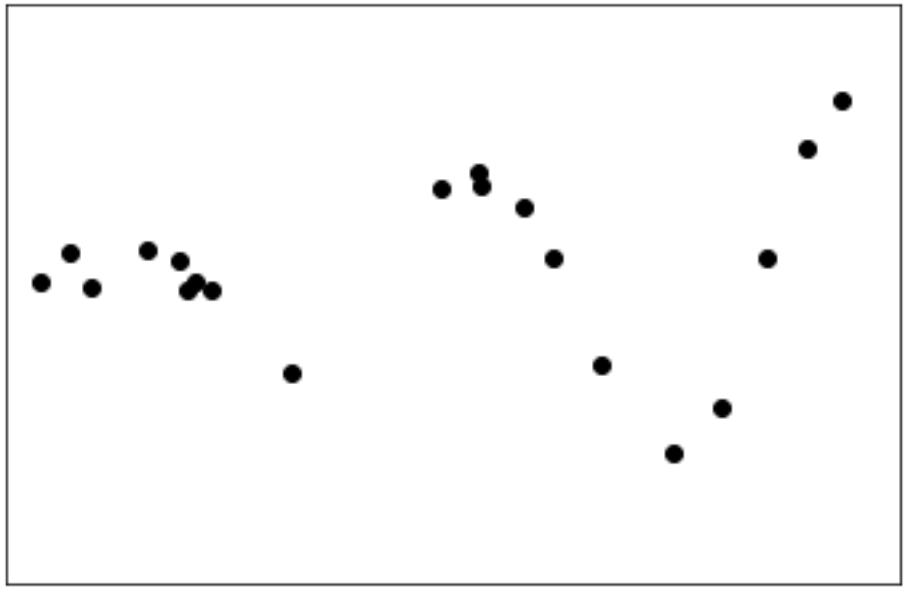
函数拟合插值 ✓

但是，高维函数的拟合并没有想象的简单

机器学习复杂性

为什么机器学习比想象的复杂?

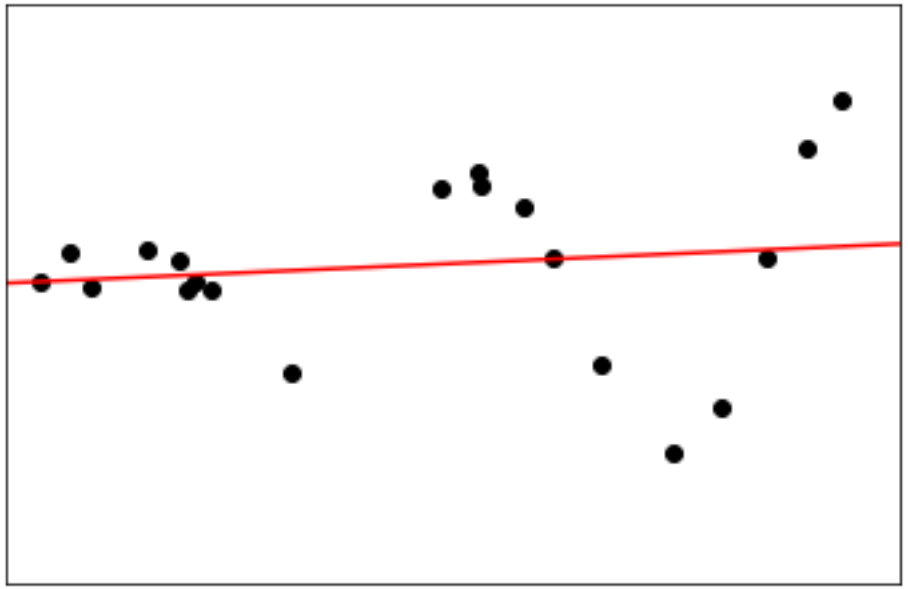
以曲线拟合为例



机器学习复杂性

为什么机器学习比想象的复杂?

以曲线拟合为例

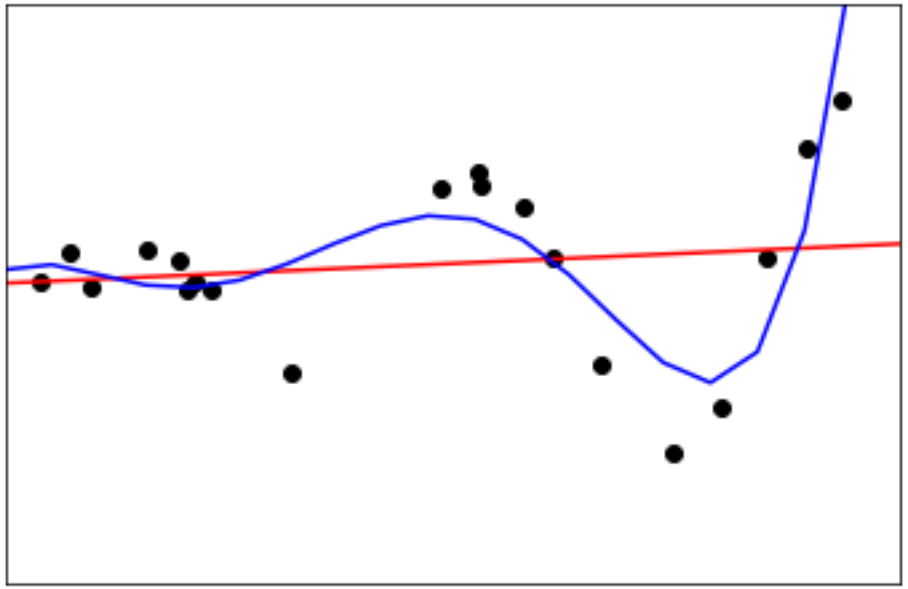


模型1

机器学习复杂性

为什么机器学习比想象的复杂?

以曲线拟合为例



模型2

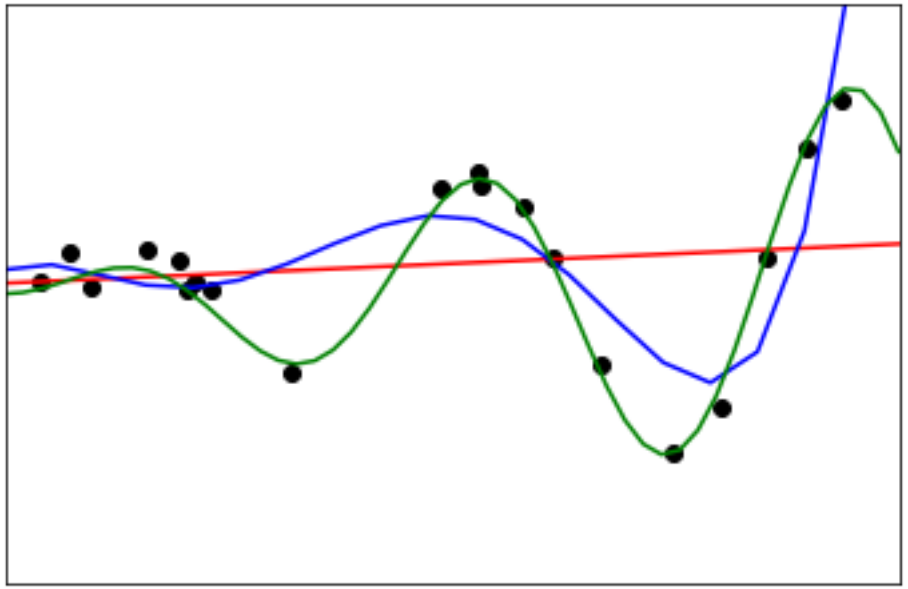
模型1

机器学习复杂性

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

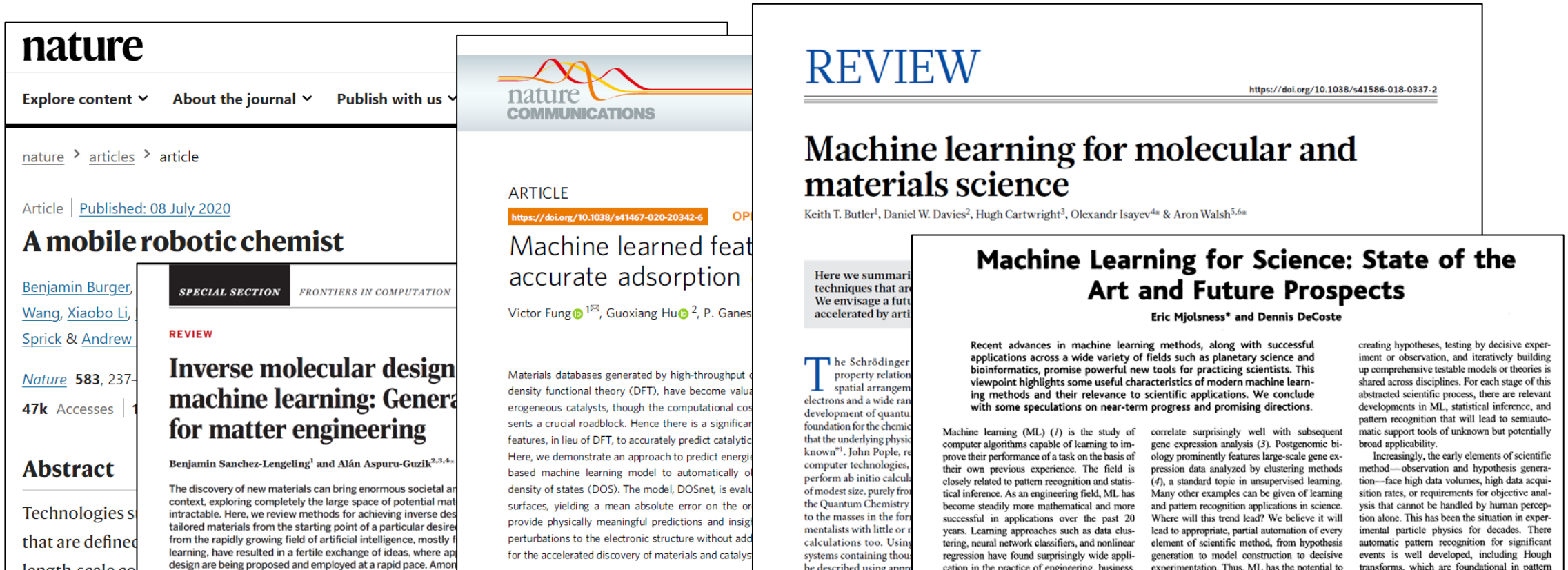
需要选择最合适的模型

目录

机器学习与化学

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–

47k Accesses

Abstract

Technologies s that are defined

SPECIAL SECTION FRONTIERS IN COMPUTATION

REVIEW

Inverse molecular design machine learning: Genera for matter engineering

Benjamin Sanchez-Lengeling[1] and Alán Aspuru-Guzik[2,3,4]

The discovery of new materials can bring enormous societal ar context, exploring completely the large space of potential mat intractable. Here, we review methods for achieving inverse des tailored materials from the starting point of a particular desire from the rapidly growing field of artificial intelligence, mostly f learning, have resulted in a fertile exchange of ideas, where ap design are being proposed and employed at a rapid pace. Amon

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-20342-6 OP

Machine learned feat accurate adsorption

Victor Fung[1], Guoxiang Hu[2], P. Ganes

Materials databases generated by high-throughput density functional theory (DFT), have become valua erogeneous catalysts, though the computational cos sents a crucial roadblock. Hence there is a significar features, in lieu of DFT, to accurately predict catalytic Here, we demonstrate an approach to predict energi based machine learning model to automatically o density of states (DOS). The model, DOSnet, is evalu surfaces, yielding a mean absolute error on the or provide physically meaningful predictions and insig perturbations to the electronic structure without add for the accelerated discovery of materials and catalys

REVIEW

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

Machine learning for molecular and materials science

Keith T. Butler[1], Daniel W. Davies[2], Hugh Cartwright[3], Olexandr Isayev[4*] & Aron Walsh[5,6*]

Here we summari techniques that ar We envisage a futu accelerated by arti

The Schrödinger property relation spatial arrangem electrons and a wide ran development of quantu foundation for the chemic that the underlying physic known[*1]. John Pople, re computer technologies, perform ab initio calcul of modest size, purely fro the Quantum Chemistry to the masses in the for mentalists with little or calculations too. Using systems containing thou be described using appr

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

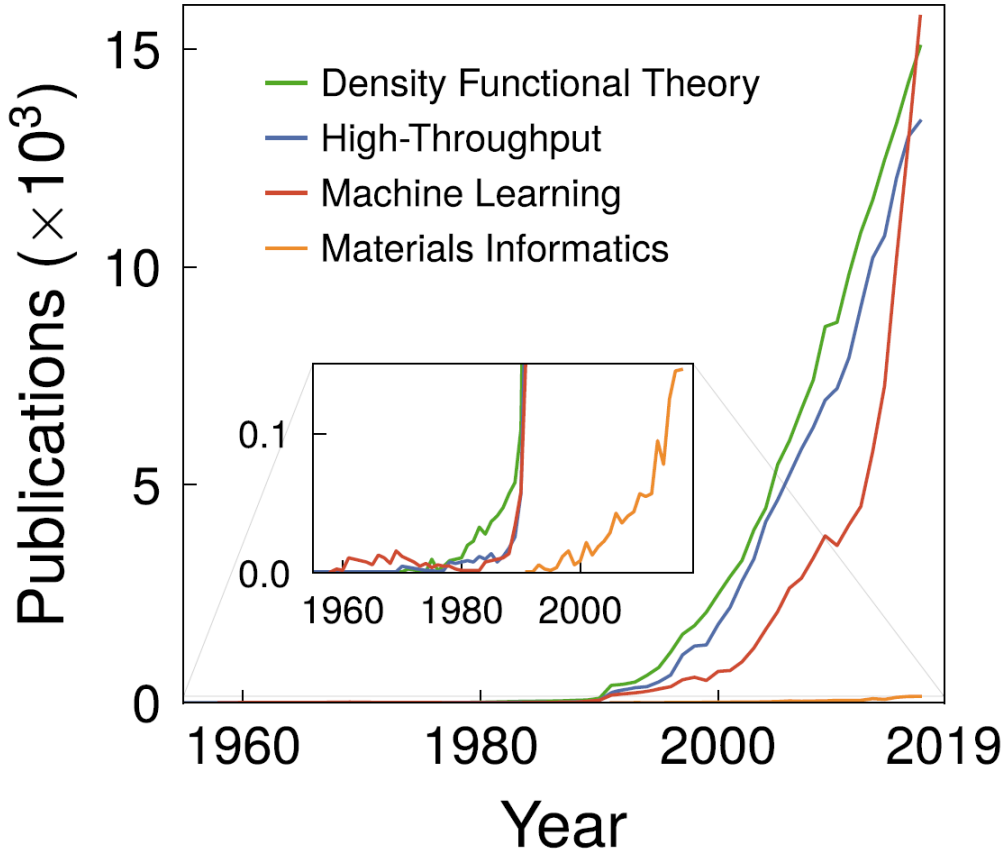
Eric Mjolsness* and Dennis DeCoste

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.

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 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

机器学习与化学



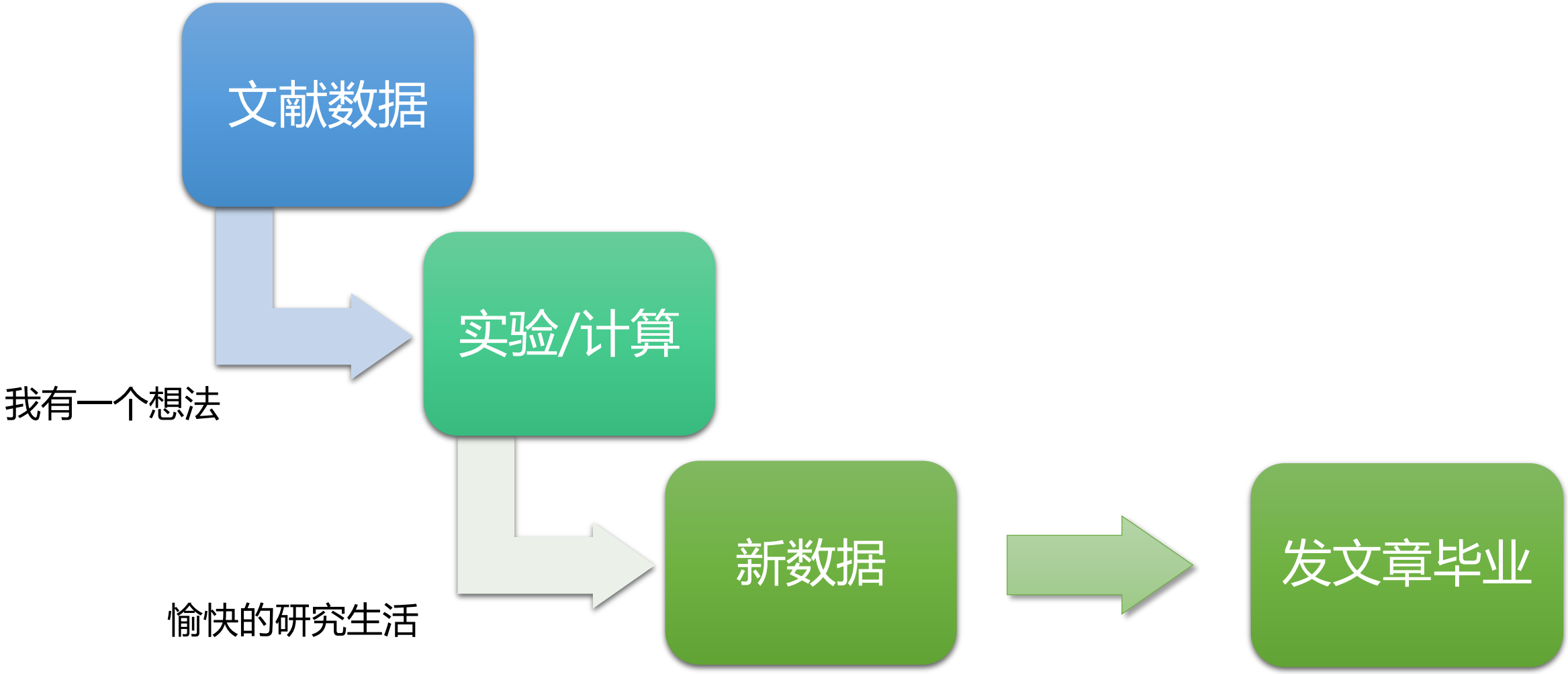
斯坦福大学2021年AI数据生成索引报告：

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%**。

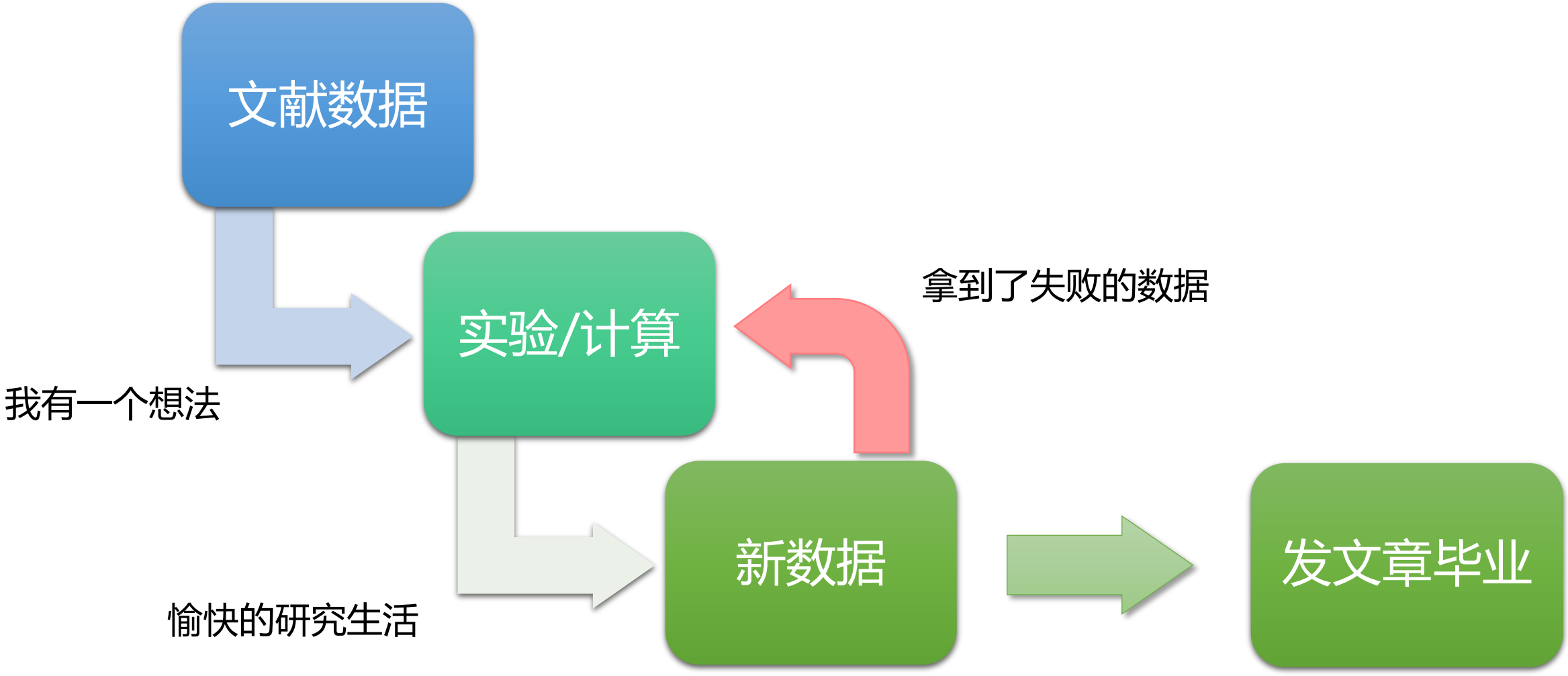
2019年，人工智能类出版物占全球所有经同行评审科学类出版物的**3.8%**，相较2011年的**1.3%**有所提升。

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

常规化学研究

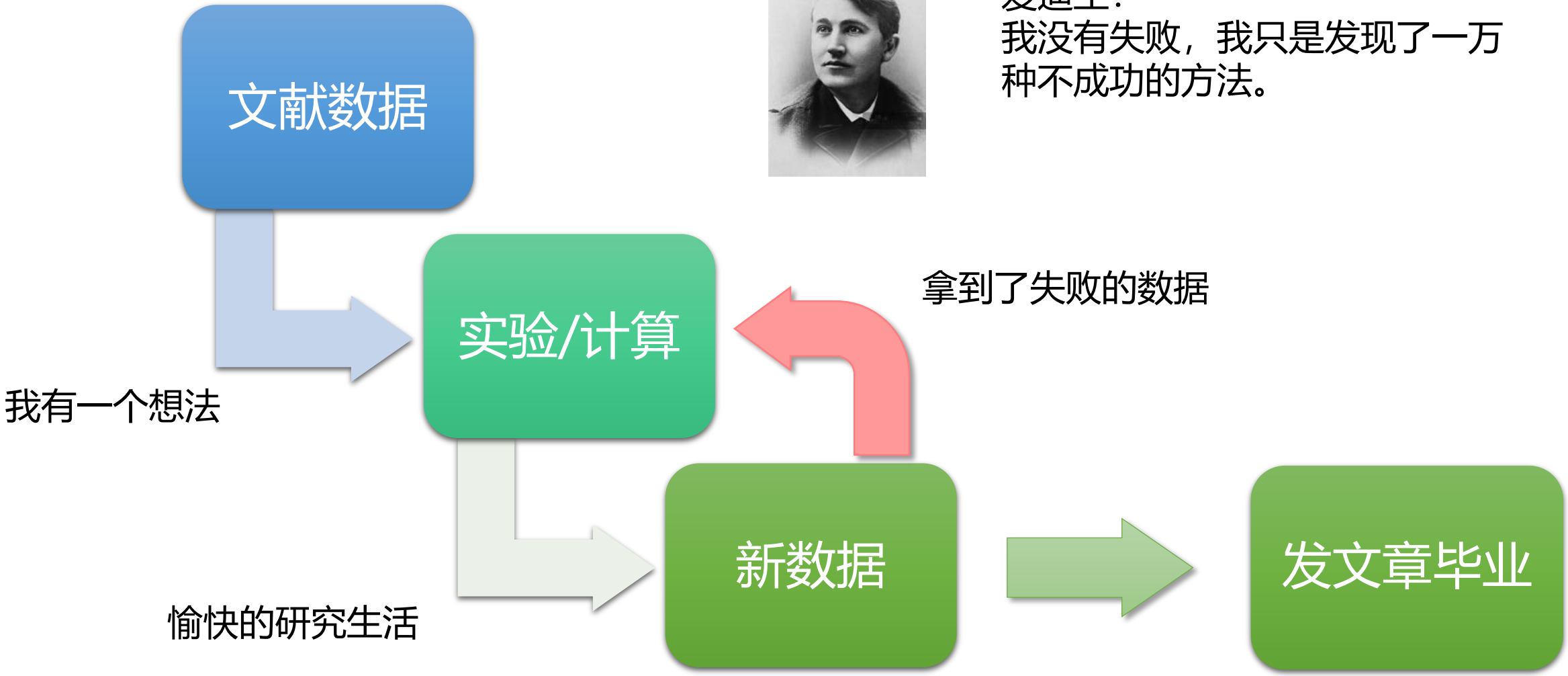


常规化学研究



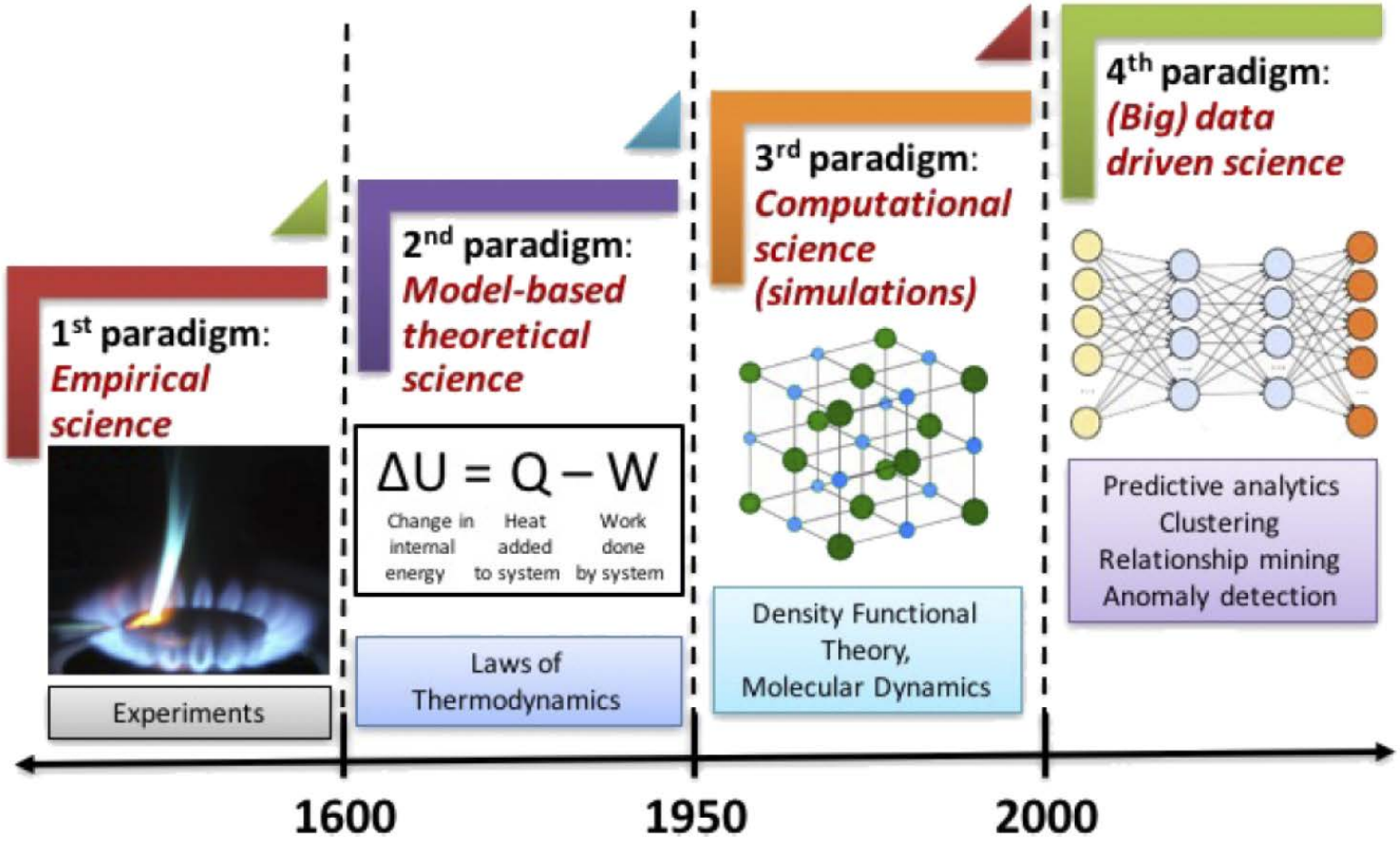
常规化学研究

爱迪生：
我没有失败，我只是发现了一万种不成功的方法。



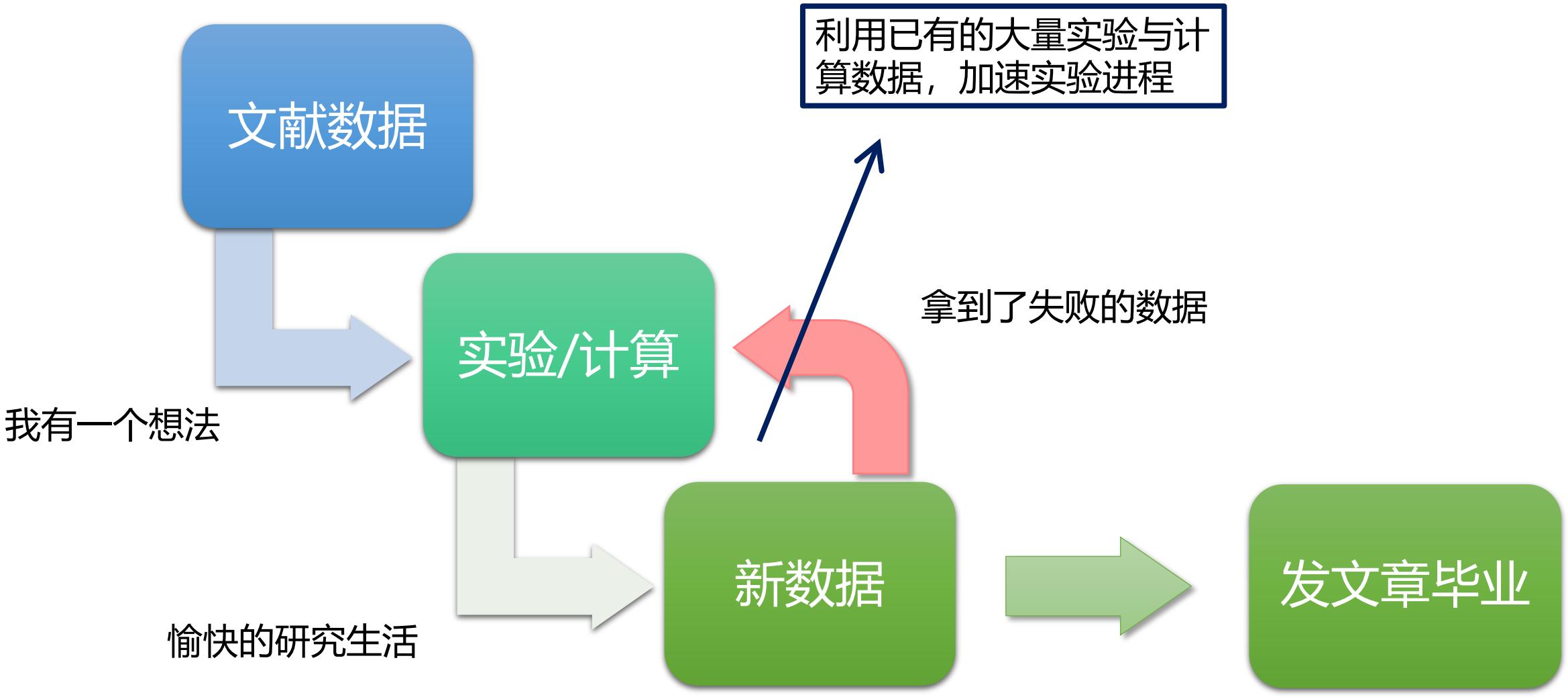
时代变了

但是，现在已经是1202年了。。。



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

数据科学时代化学研究


利用已有的大量实验与计算数据，加速实验进程
文献数据
实验/计算
拿到了失败的数据
我有一个想法
新数据
发文章毕业
愉快的研究生活

目录

Python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

scikit-learn

Machine Learning in Python

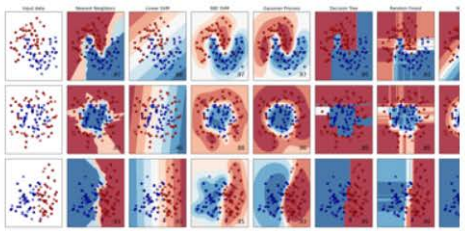
Getting Started | Release Highlights for 1.0 | GitHub

- 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...

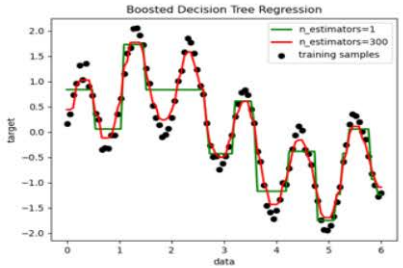


Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: SVR, nearest neighbors, random forest, and more...

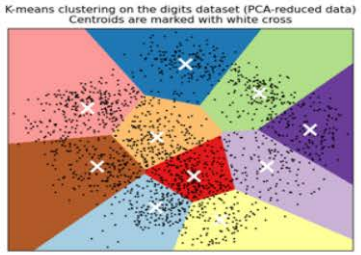


Examples

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.