第三讲 连接主义与机器学习

郑子杰 韩思瑶
北京市十一学校

所有课件可以用来教学使用，未经允许不得以盈利为目进行分享

目录 Content

- 人工智能的三大流派
- 我们这节课要学习的人工智能——机器学习 Machine learning
- 机器学习的基本流程

人工智能的三大流派

- 符号主义 Symbolicism
  - 用计算机进行数学证明和推导；例如Mathematica

- 行为主义 Actionism
  - 会动的（机器人）才配叫人工智能

- 连接主义 Connectionism
  - 以神经网络为代表的、基于大量数据生成数学模型并进行预测的人工智能实现过程，计算机基于数据建立模型的流程、范式和算法一般被叫作机器学习（Machine Learning），也是当下的**主流**

人工智能的三大流派

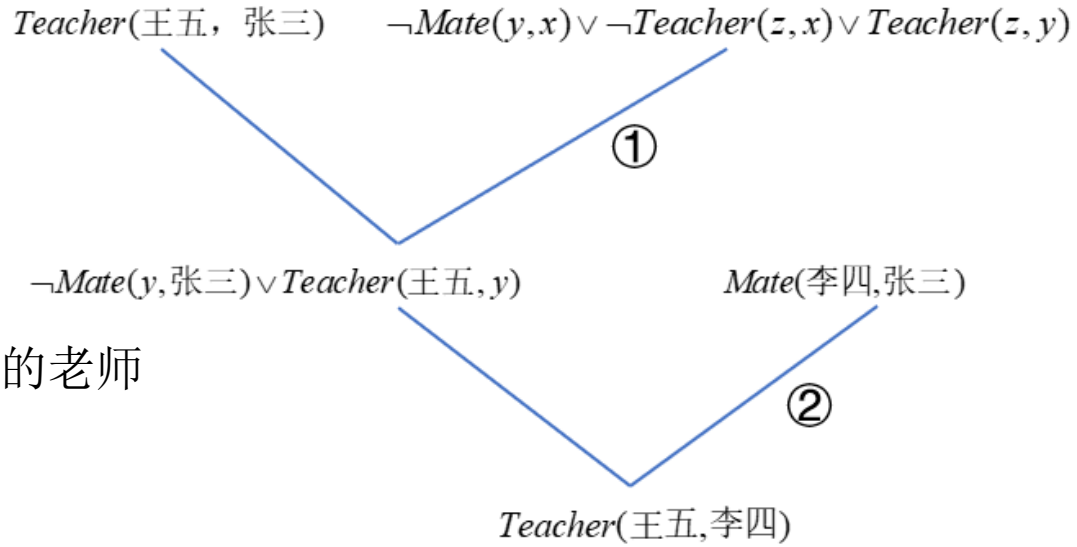
- 符号主义 Symbolicism
  - **用计算机进行数学证明和逻辑推理**

p_1：王五是张三的老师

p_2：张三和李四是同班同学

p_3：如果x和y是一个班的，则x的老师也一定是y的老师

现在的问题是：李四的老师是谁？



人工智能的三大流派

- 行为主义 Actionism
  - 会动的（机器人）才配叫人工智能

人工智能的三大流派

- 符号主义 Symbolicism
  - 用计算机进行数学证明和推导；例如Mathematica

- 行为主义 Actionism
  - 会动的（机器人）才配叫人工智能

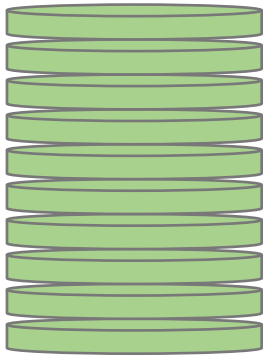
- 连接主义 Connectionism
  - 以神经网络为代表的、基于大量数据生成数学模型并进行预测的人工智能实现过程，计算机基于数据建立模型的流程、范式和算法一般被叫作机器学习（Machine Learning），也是当下的**主流**

目录 Content

- 人工智能的三大流派
- 我们这节课要学习的人工智能——机器学习 Machine learning
- 机器学习的基本流程

机器学习 Machine Learning

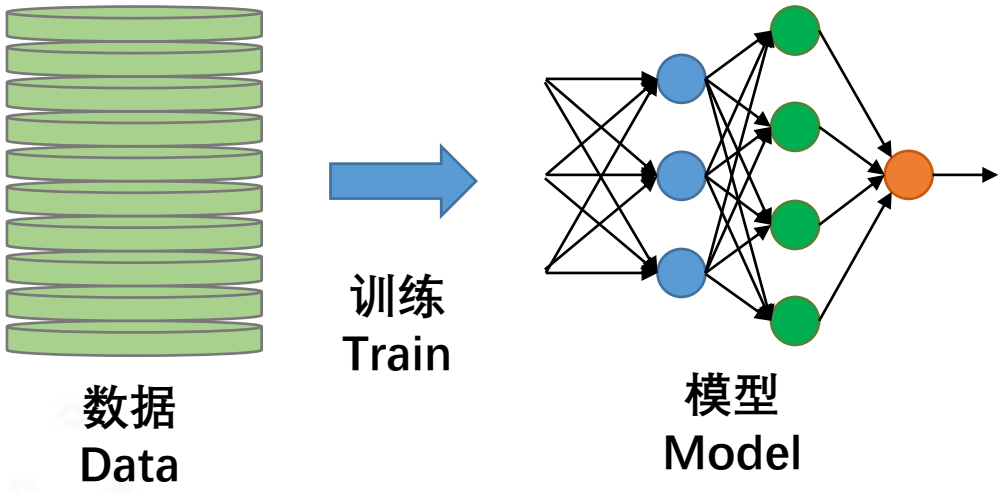
- **机器学习（连接主义人工智能）**
- **Machine Learning**
- 基于**大量数据**训练**数学模型**并在新的数据上**进行预测（测试）**的人工智能实现过程



数据
Data

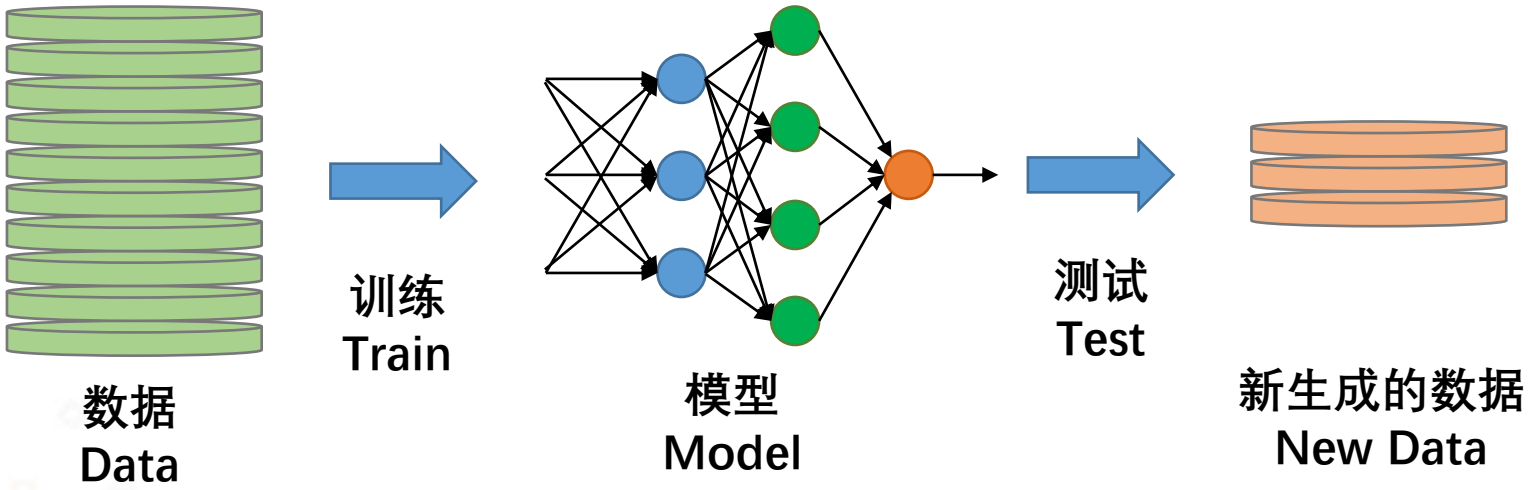
机器学习 Machine Learning

- **机器学习（连接主义人工智能）**
- **Machine Learning**
- 基于**大量数据**训练**数学模型**并在新的数据上**进行预测（测试）**的人工智能实现过程



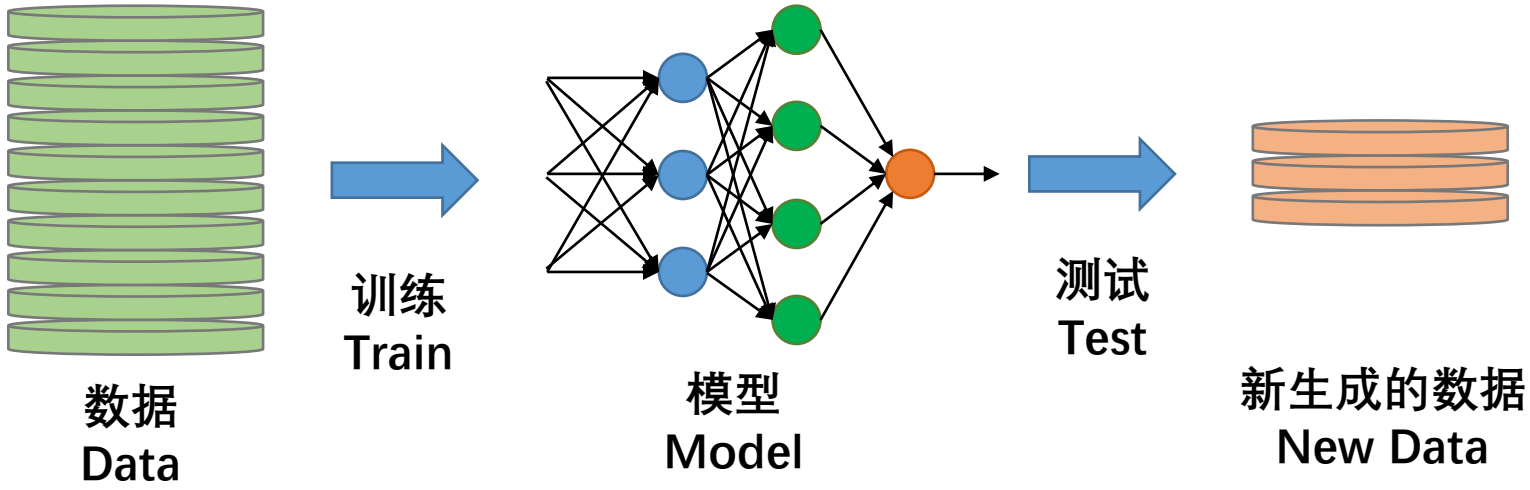
机器学习 Machine Learning

- **机器学习（连接主义人工智能）**
- **Machine Learning**
- 基于**大量数据**训练**数学模型**并在新的数据上**进行预测（测试）**的人工智能实现过程



机器学习 Machine Learning

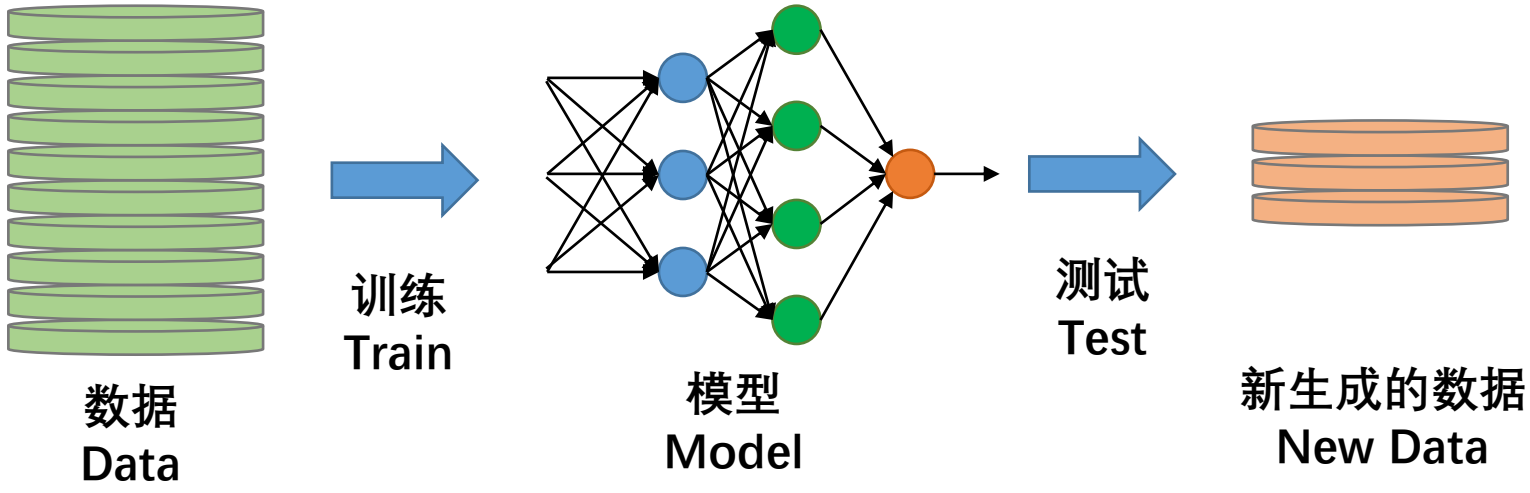
- 机器学习（例） Example
- 碑文修复 Rubbing Restoration using Machine Learning

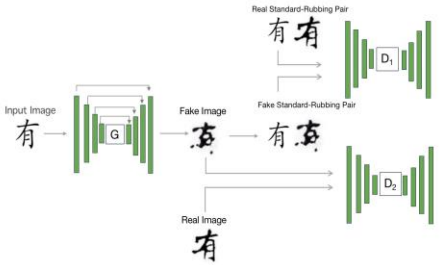


本页案例来源：Gongbo Sun, Zijie Zheng, Ming Zhang, “End-to-End Rubbing Restoration Using Generative Adversarial Networks——Taking Zhang Menglong Stela As An Example“

机器学习 Machine Learning

- 机器学习（例） Example
- 碑文修复 Rubbing Restoration using Machine Learning

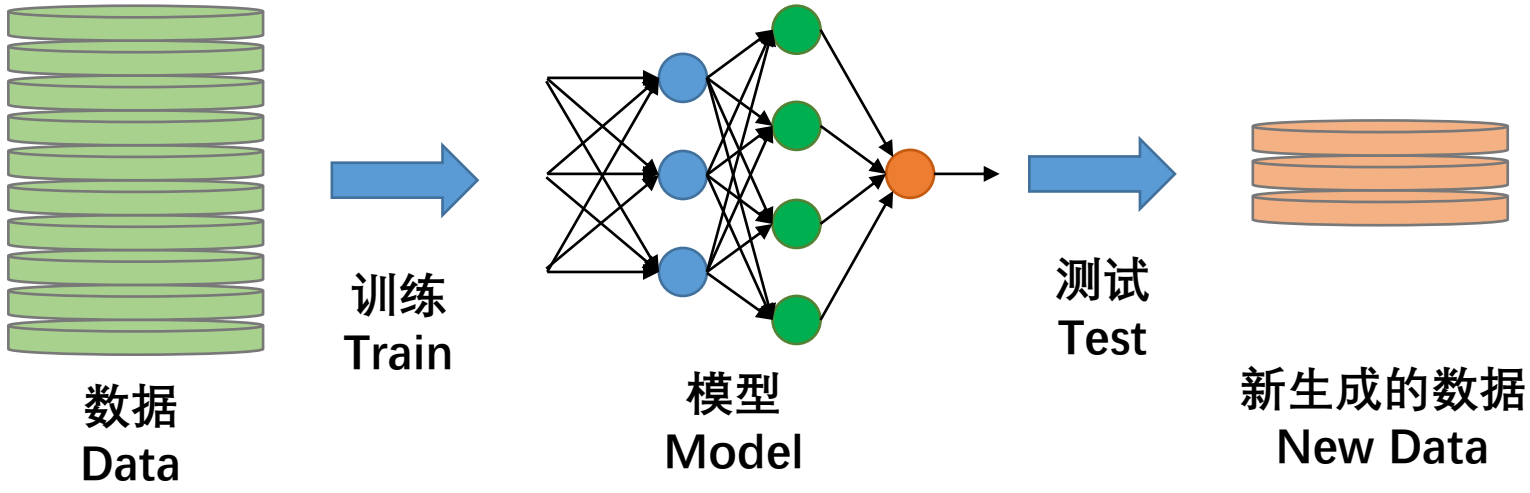


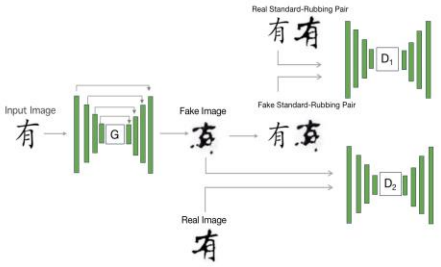


本页案例来源：Gongbo Sun, Zijie Zheng, Ming Zhang, “End-to-End Rubbing Restoration Using Generative Adversarial Networks——Taking Zhang Menglong Stela As An Example“

机器学习 Machine Learning

- 机器学习（例） Example
- 碑文修复 Rubbing Restoration using Machine Learning





修复碑文
Rubbing
Restoration

本页案例来源：Gongbo Sun, Zijie Zheng, Ming Zhang, “End-to-End Rubbing Restoration Using Generative Adversarial Networks——Taking Zhang Menglong Stela As An Example“

机器学习 Machine Learning

- 为什么这样的人工智能实现过程合理?

人的智能 Intelligence of Human Beings	人工智能 Artificial Intelligence

机器学习 Machine Learning

- 为什么这样的人工智能实现过程合理?

人的智能 Intelligence of Human Beings	人工智能 Artificial Intelligence
① 不断地学习知识	

机器学习 Machine Learning

- 为什么这样的人工智能实现过程合理?

人的智能 Intelligence of Human Beings	人工智能 Artificial Intelligence
① 不断地学习知识	
② 综合所学知识训练解决各类问题的基本技能	

机器学习 Machine Learning

- 为什么这样的人工智能实现过程合理?

人的智能 Intelligence of Human Beings	人工智能 Artificial Intelligence
① 不断地学习知识	
② 综合所学知识训练解决各类问题的基本技能	
③ 使用训练好的技能解决类似的问题	

机器学习 Machine Learning

- 为什么这样的人工智能实现过程合理?

人的智能 Intelligence of Human Beings	人工智能 Artificial Intelligence
① 不断地学习知识	① 积累大量数据
② 综合所学知识训练解决各类问题的基本技能	
③ 使用训练好的技能解决类似的问题	

机器学习 Machine Learning

- 为什么这样的人工智能实现过程合理?

人的智能 Intelligence of Human Beings	人工智能 Artificial Intelligence
① 不断地学习知识	① 积累大量数据
② 综合所学知识训练解决各类问题的基本技能	② 根据数据训练解决特定问题的数学模型
③ 使用训练好的技能解决类似的问题	

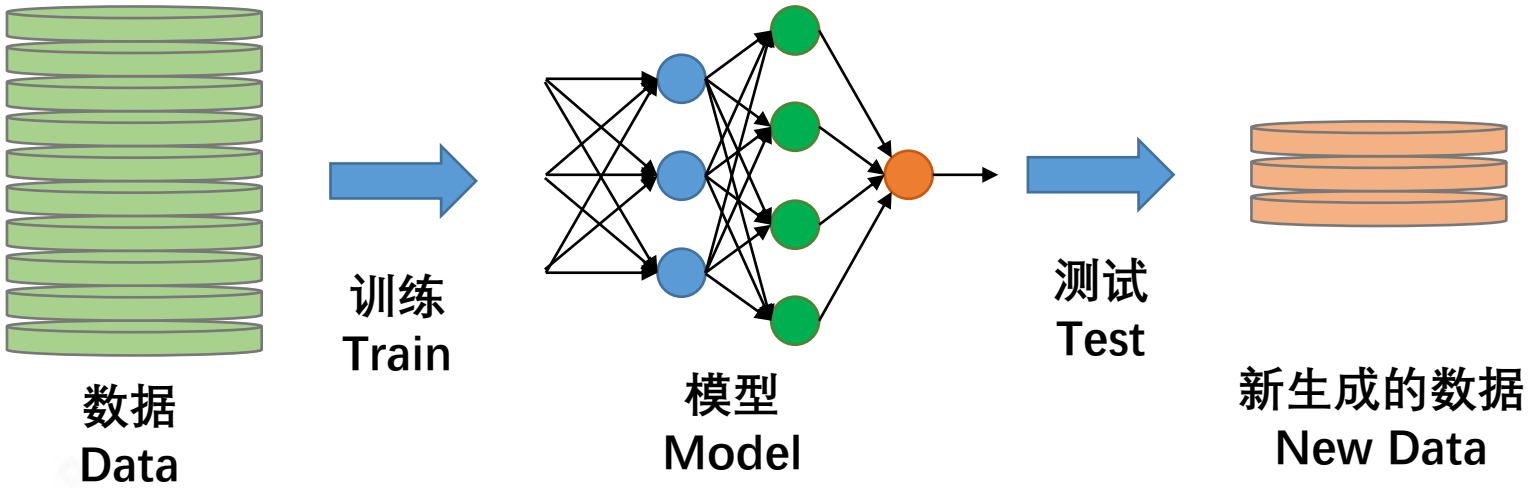
机器学习 Machine Learning

- 为什么这样的人工智能实现过程合理?

人的智能 Intelligence of Human Beings	人工智能 Artificial Intelligence
① 不断地学习知识	① 积累大量数据
② 综合所学知识训练解决各类问题的基本技能	② 根据数据训练解决特定问题的数学模型
③ 使用训练好的技能解决类似的问题	③ 使用数学模型处理特定问题中生成的新的数据

机器学习 Machine Learning

- 按理说这个想法很自然，人们之前不是这么处理问题的么？



机器学习 Machine Learning

- 过去 In 20th century
  - ①收集数据的流程繁琐
  - ②存储数据的能力匮乏
  - ③分析数据的算力也不够

机器学习 Machine Learning

- 过去 In 20th century
  - ①收集数据的流程繁琐
  - ②存储数据的能力匮乏
  - ③分析数据的算力也不够

  - 只能基于少量数据分析和建立模型

机器学习 Machine Learning

- 过去 In 20th century
  - ①收集数据的流程繁琐
  - ②存储数据的能力匮乏
  - ③分析数据的算力也不够

  - 只能基于少量数据分析和建立模型

- 如何基于少量数据生成模型?

基于少量数据生成模型
Modeling on Small Amount of Data

- 例如：一个年级有300人，只统计了30个人的学习情况，如何估计全年级的学习情况？

基于少量数据生成模型
Modeling on Small Amount of Data

- **例如：一个年级有300人，只统计了30个人的学习情况，如何估计全年级的学习情况？**

- **覆盖各类学生**：这30个人肯定得从这300人的一个总体里面抽，而且尽可能的**均匀。**

基于少量数据生成模型
Modeling on Small Amount of Data

- **例如：一个年级有300人，只统计了30个人的学习情况，如何估计全年级的学习情况？**

- **个体分析不靠谱**：因为抽的太少，**无论**建立什么样的模型，怎么分析这30个人的成绩与学习习惯、学习时间等因素，去分析剩下270个人中任何一个个体，都不是那么靠谱和自信，因为270个人中出现例外的可能性太大了。

基于少量数据生成模型
Modeling on Small Amount of Data

- **例如：一个年级有300人，只统计了30个人的学习情况，如何估计全年级的学习情况？**

- **群体特征分析合理性**：对于一些事情，或许可以做的很好，例如如果抽样抽的好的话，用这30个人平均值，去估计总体300人的平均值。

基于少量数据生成模型
Modeling on Small Amount of Data

- 例如：一个年级有300人，只统计了30个人的学习情况，如何估计全年级的学习情况？

- 从统计学角度看
- 覆盖各类学生
  - 随机抽样 randomly sampling，独立同分布 independently identically distribution, i.i.d

基于少量数据生成模型
Modeling on Small Amount of Data

- 例如：一个年级有300人，只统计了30个人的学习情况，如何估计全年级的学习情况？

- 从统计学角度看
- 覆盖各类学生
  - 随机抽样 randomly sampling，独立同分布 independently identically distribution, i.i.d
- 个体分析不靠谱
  - 统计学不关注个体的分析

基于少量数据生成模型
Modeling on Small Amount of Data

- 例如：一个年级有300人，只统计了30个人的学习情况，如何估计全年级的学习情况？

- 从统计学角度看
- 覆盖各类学生
  - 随机抽样 randomly sampling，独立同分布 independently identically distribution, i.i.d
- 个体分析不靠谱
  - 统计学不关注个体的分析
- 群体特征分析合理性
  - 统计学关注群体分析
    统计量（例如平均值、方差），用样本估计总体，大数定律与中心极限定理…

- 现在，时代变了：**积累大量数据这件事不再是难事**

机器学习 Machine Learning

- 我们现在面对的问题大部分是这样的
- 例如：一个年级有300人，我们统计了290个人（而非30人）的学习情况，如何估计全年级的学习情况？

机器学习 Machine Learning

- 我们现在面对的问题大部分是这样的
- 例如：一个年级有300人，我们统计了290个人（而非30人）的学习情况，如何估计全年级的学习情况？

- 群体分析无现实意义
- 就剩10个人了，自然每个个体的分析都很重要

机器学习 Machine Learning

- **我们现在面对的问题大部分是这样的**
- **例如：一个年级有300人，我们统计了290个人（而非30人）的学习情况，如何估计全年级的学习情况？**

- **群体分析无现实意义**
- **就剩10个人了，自然每个个体的分析都很重要**

- **我们更关注个体分析**
- 因为样本量足够多（接近总体），大部分情况都涵盖了，所以用这290个人的成绩和学习习惯等因素建立模型会很靠谱，然后去预测这个10个个体每个人的成绩

统计学 vs 机器学习
Statistics vs Machine Learning

- 传统统计学——更关注群体分析，用满足独立同分布的少量样本的统计量（例如均值）可以估计总体的统计量，而且随着样本量的增大，会越来越准

统计学 vs 机器学习
Statistics vs Machine Learning

- 传统统计学——更关注群体分析，用满足独立同分布的少量样本的统计量（例如均值）可以估计总体的统计量，而且随着样本量的增大，会越来越准

- 机器学习——更关注个体预测：在数据量足够大的情况下，暨拿到了接近总体数据量的大量数据，我们只需要关心如何建立模型，准确地预测剩下的/新产生的每一个个体

统计学 vs 机器学习
Statistics vs Machine Learning

	统计学 Statictics	机器学习 Machine Learning
使用的前提不同	数据是少量样本，用样本估计总体	收集到的数据已经很接近于总体
常用方法	**群体分析** ① 统计量分析（均值、方差、中位数、四分位数、矩…） ② 相关分析（变量之间是否存在相关关系） ③ 回归分析（变量之间的函数关系） **④ 假设检验（分析的可信度的数学检验）**	**个体预测** ① 训练：使用已知数据**建立模型** ② 预测：使用新数据验证模型

统计学 vs 机器学习
Statistics vs Machine Learning

- 多大的样本量算是少量样本？多大的样本量属于大量样本（接近总体的样本量）？

统计学 vs 机器学习
Statistics vs Machine Learning

- 这并不是个数学问题，什么叫做大量数据，**主要取决于领域本身的约定**：

- 例如：某种罕见病，全国只有不到1万个患者，收集到了8000个数据，完全可以叫大量数据.

统计学 vs 机器学习
Statistics vs Machine Learning

- 这并不是个数学问题，什么叫做大量数据，**主要取决于领域本身的约定**：

- 例如：某种罕见病，全国只有不到1万个患者，收集到了8000个数据，完全可以叫大量数据.

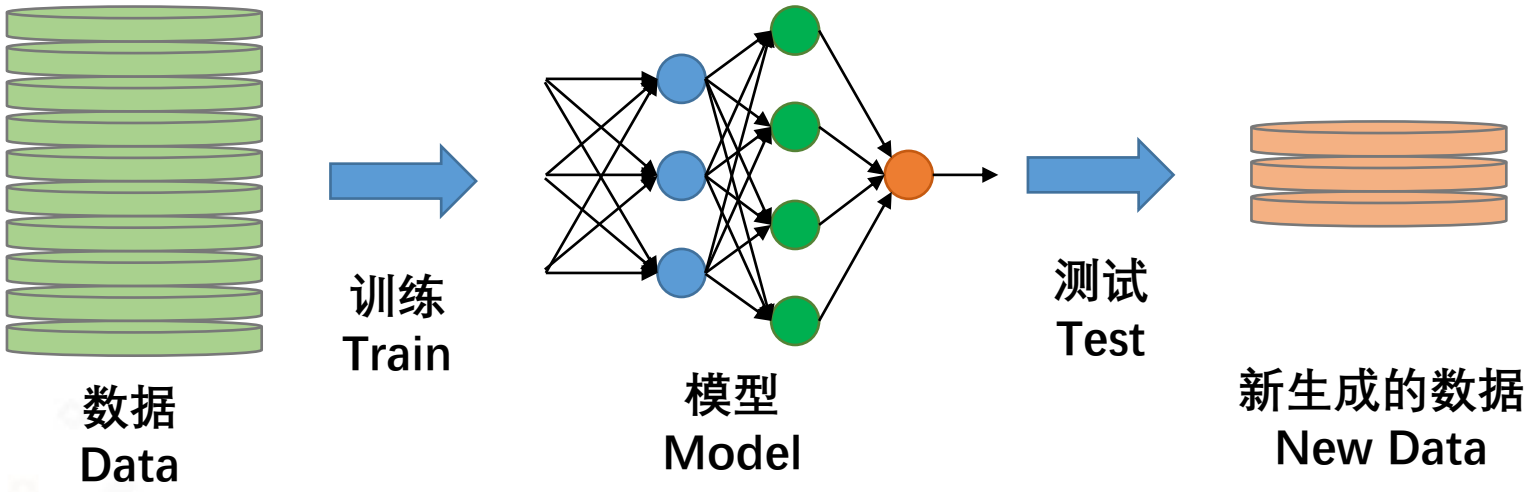
- 对比：现在一个图片数据集动辄就是几千万张图片.

目录 Content

- 人工智能的三大流派
- 我们这节课要学习的人工智能——机器学习 Machine learning
- 机器学习的基本流程

机器学习的基本流程 Process of Machine Learning

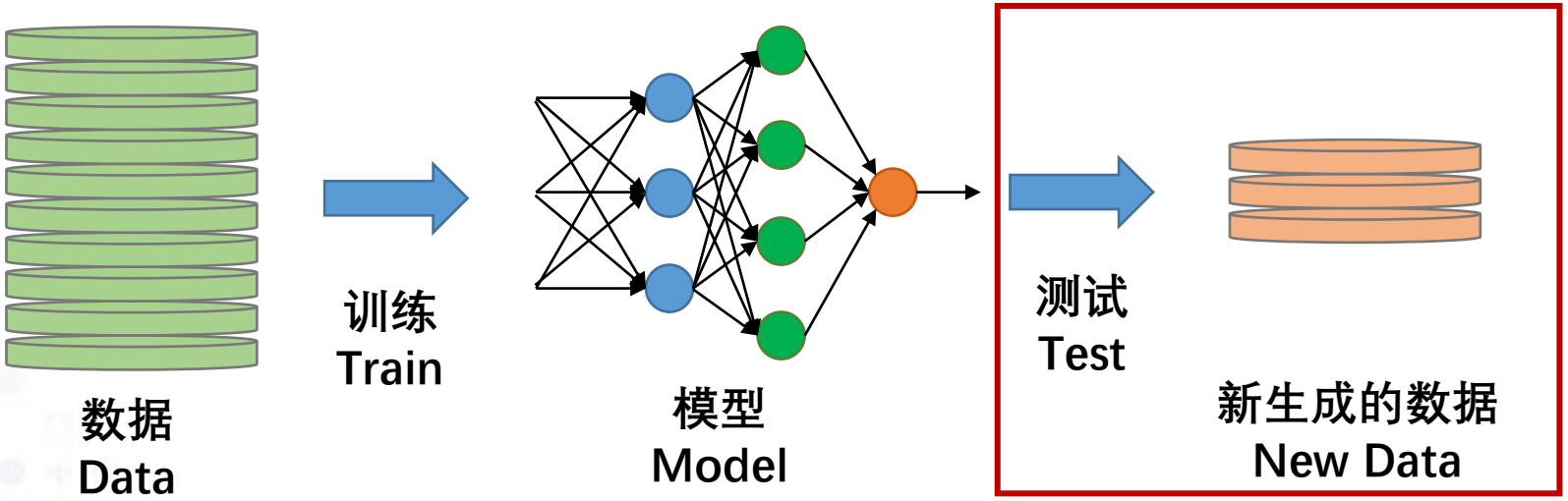
- **机器学习（连接主义人工智能）**
- **Machine Learning**
- 基于**大量数据**训练**数学模型**并在新的数据上**进行预测（测试）**的人工智能实现过程



机器学习的基本流程 Process of Machine Learning

- 新数据之殇

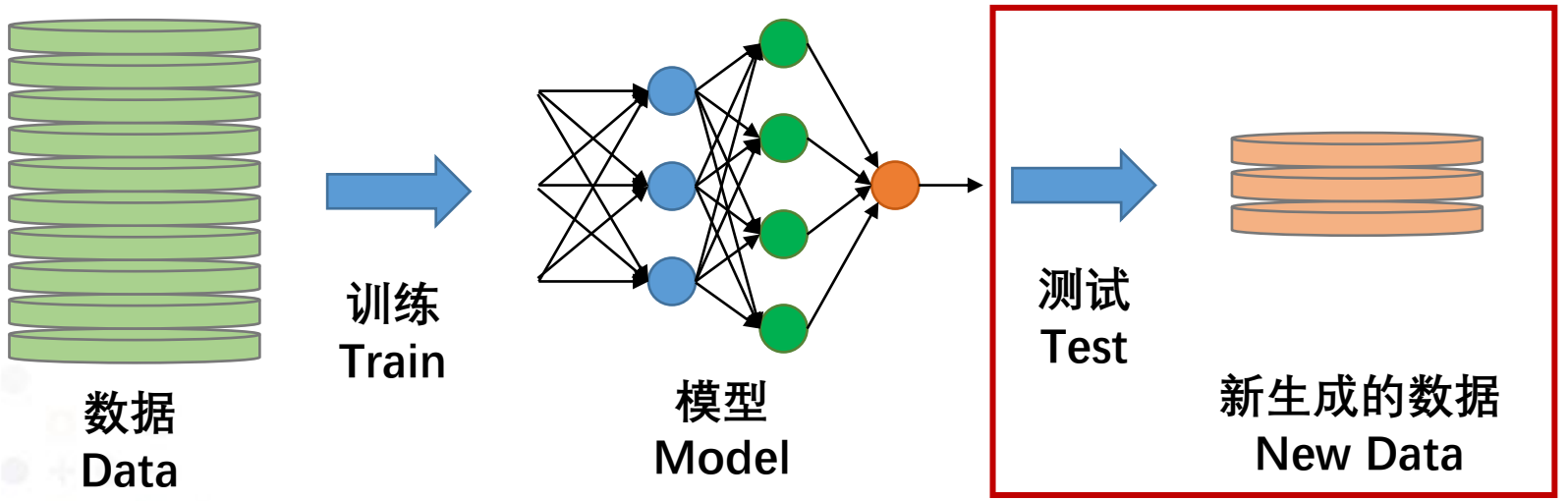
- “新”意味着用的时候刚产生，既然是刚产生，那么在“新”产生之前没人知道模型的好坏。这显然是不能接受的。



机器学习的基本流程 Process of Machine Learning

- 新数据之殇

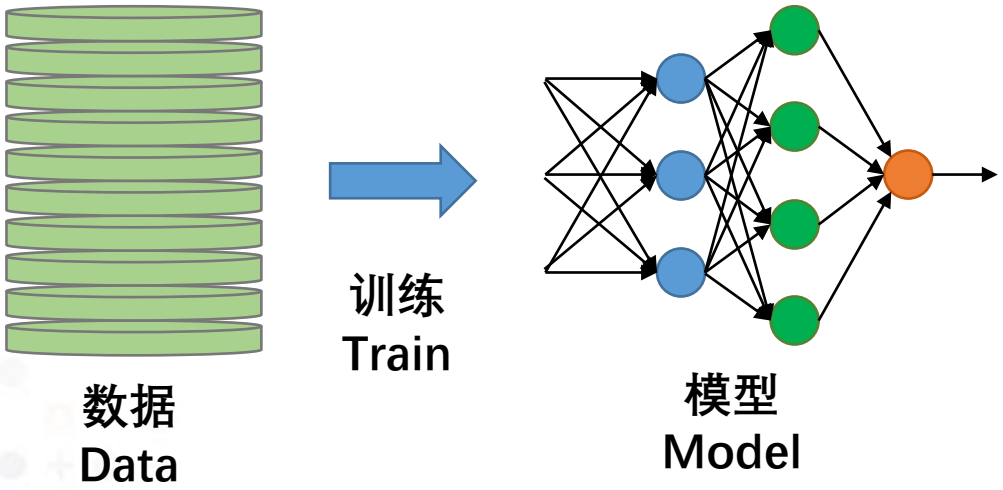
- 例如：建立了模型预测台风，结果台风来了才知道这个模型是否有效，那要这个模型的意义何在。



机器学习的基本流程 Process of Machine Learning

- 新数据之殇

- 结论：在我们建立模型时，是看不到新数据的，也就不知道模型建的好不好



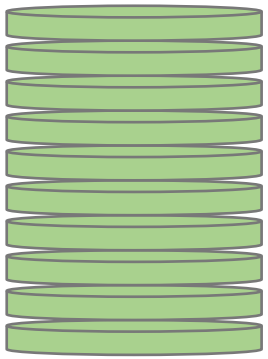
机器学习的基本流程 Process of Machine Learning

- 解决新数据之殇 Deal with the problem of new data

- 用少部分老数据假装新数据，来测试模型好坏！

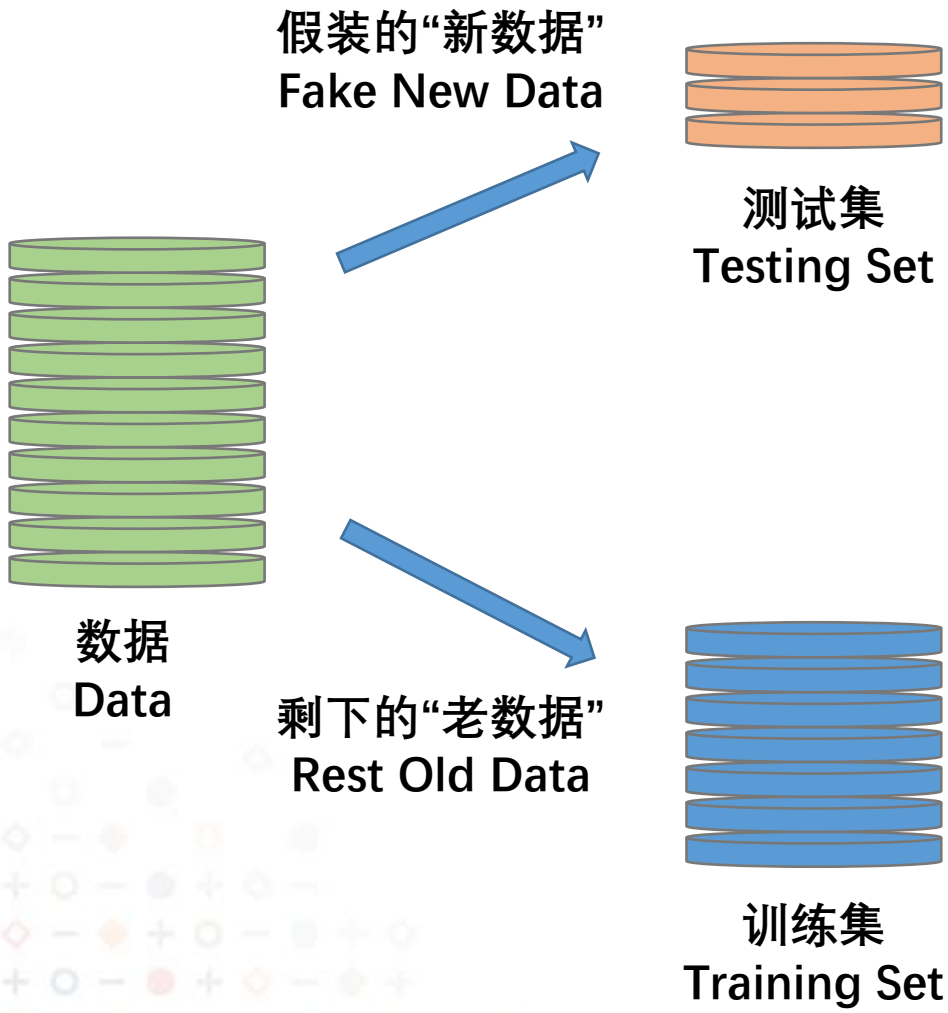
- 具体而言，在已经收集的数据中，随机抽取一部分当作“新”数据，称为测试数据（Testing Set），而剩下的那些当作“老”数据称为训练数据（Training Set）用于生成模型。
- 一般“新”、“老”比例可以为1:9或者2:8或者3:7等等。

机器学习的基本流程 Process of Machine Learning

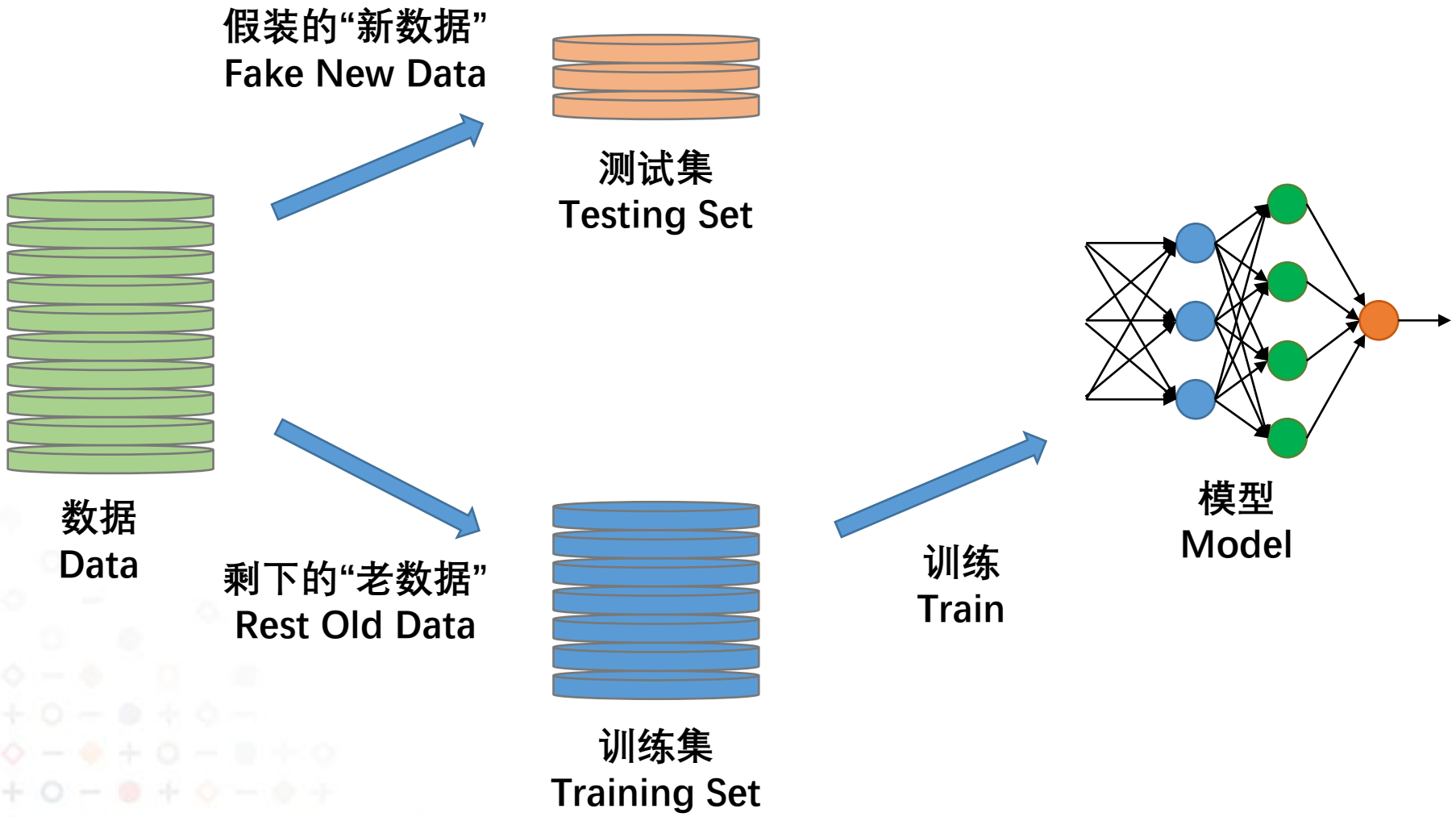


数据
Data

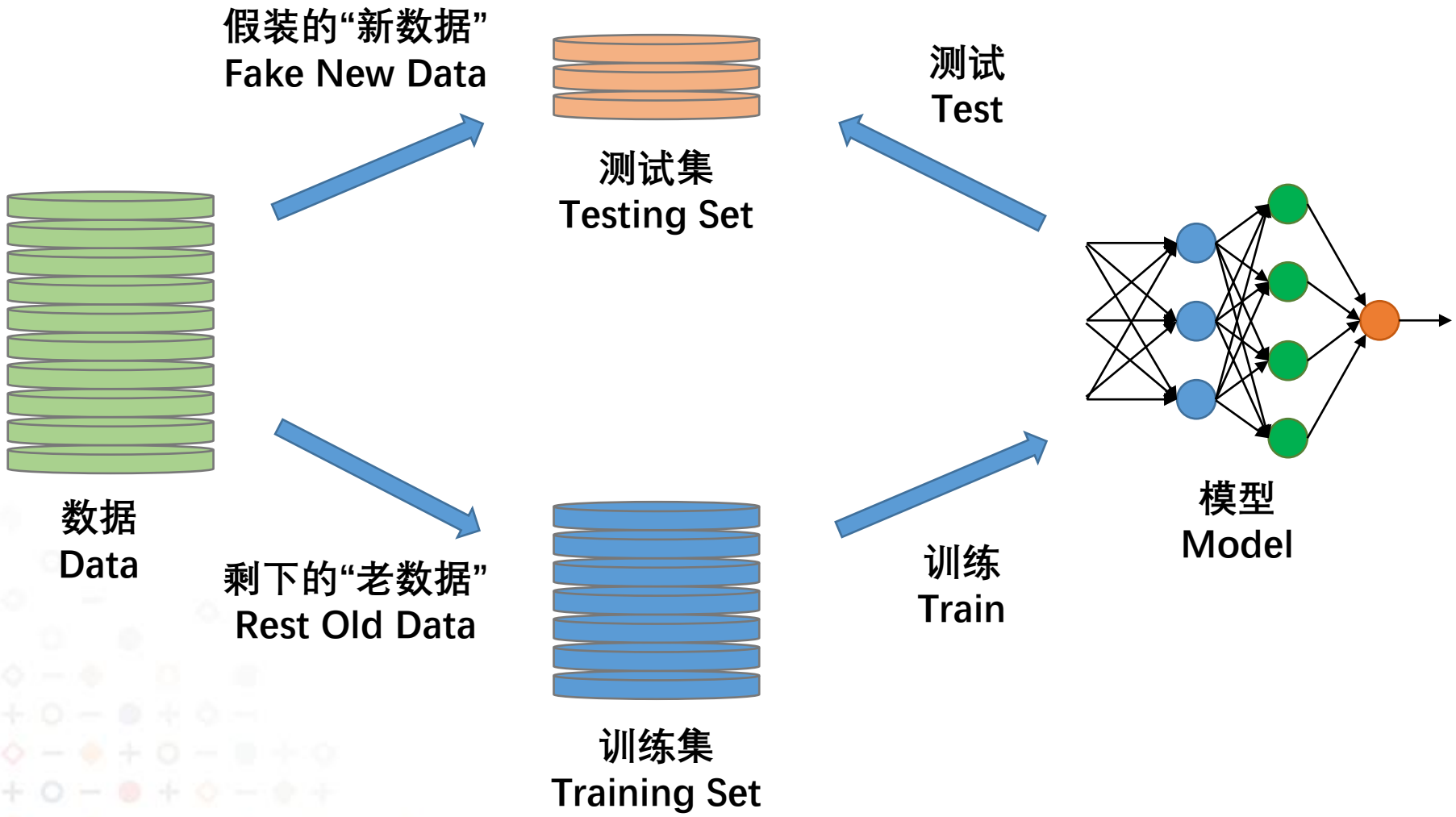
机器学习的基本流程 Process of Machine Learning



机器学习的基本流程 Process of Machine Learning



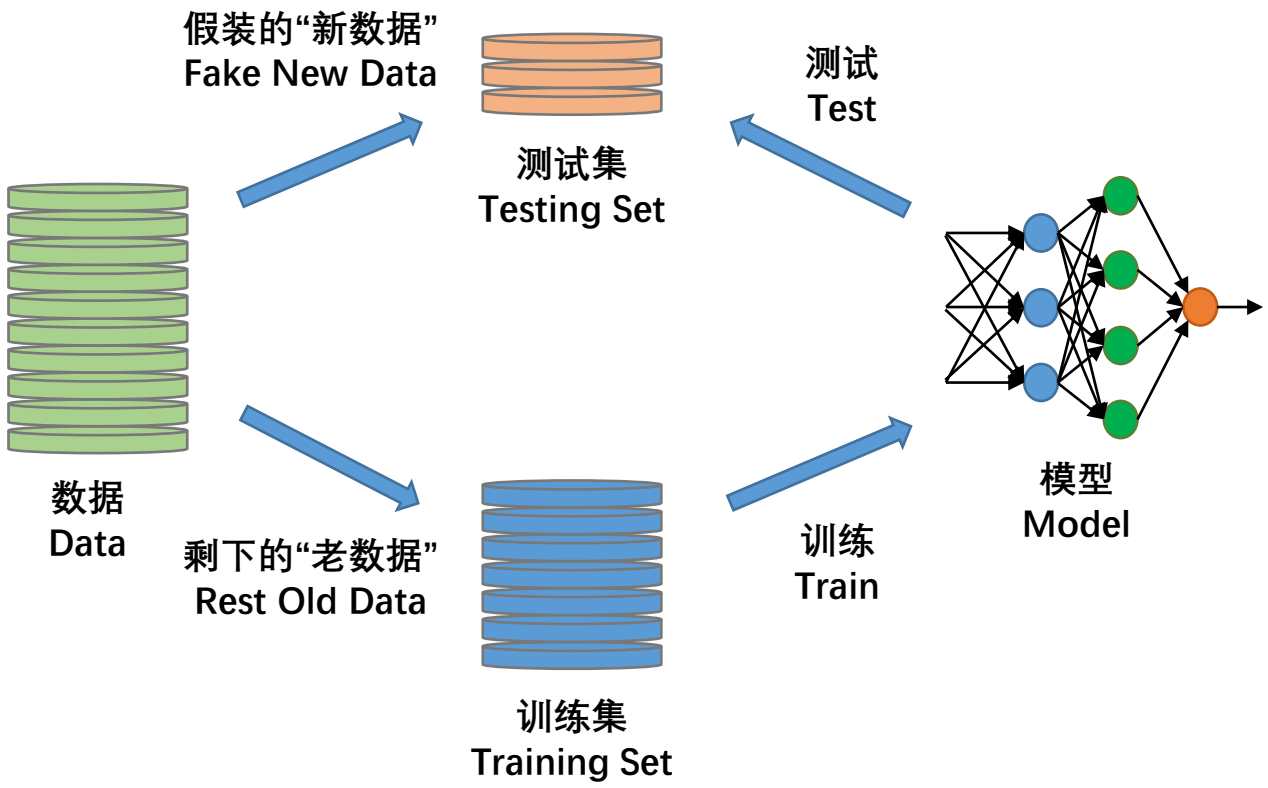
机器学习的基本流程 Process of Machine Learning



机器学习的基本流程 Process of Machine Learning

- **注意事项！Warnings!**

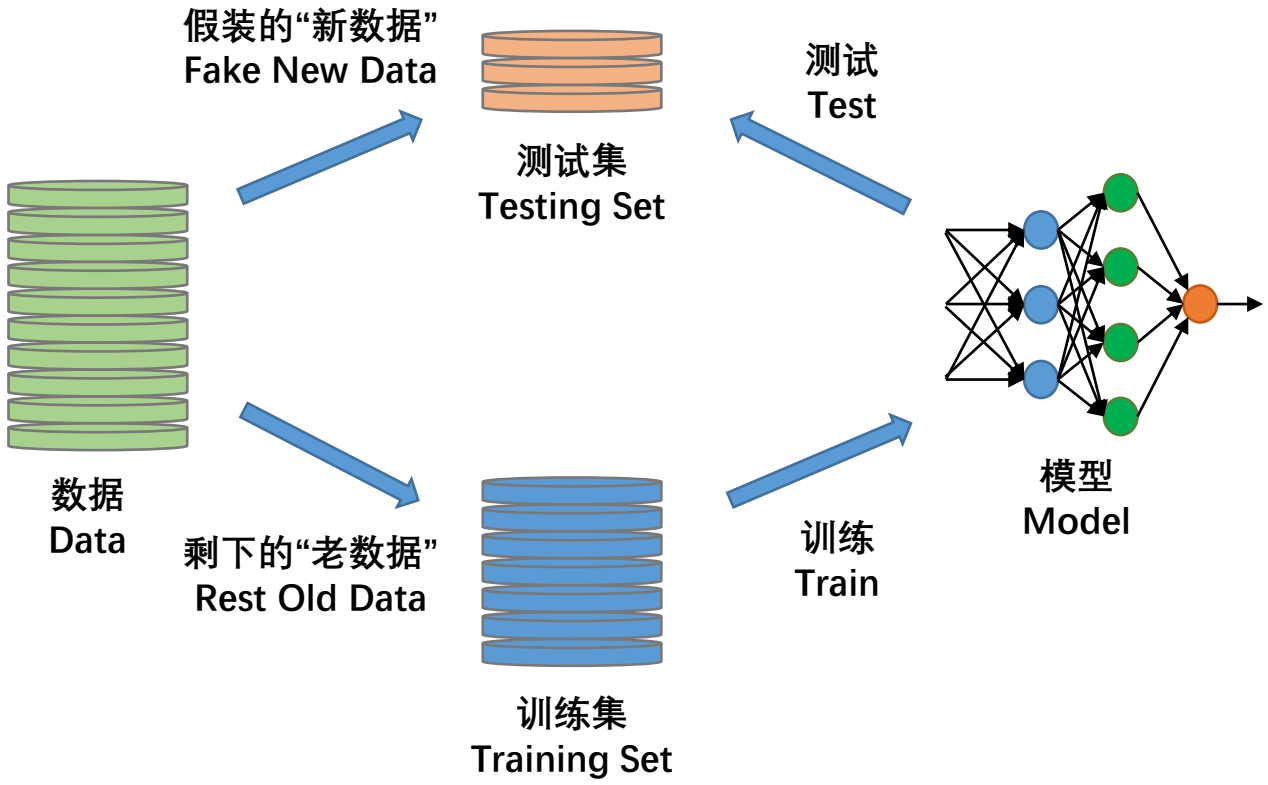
- **① 训练集的数据量一定要远大于测试集。这样才能符合机器学习的做事逻辑，用大量数据建立的模型去预测新数据。**



机器学习的基本流程 Process of Machine Learning

- **注意事项！Warnings!**

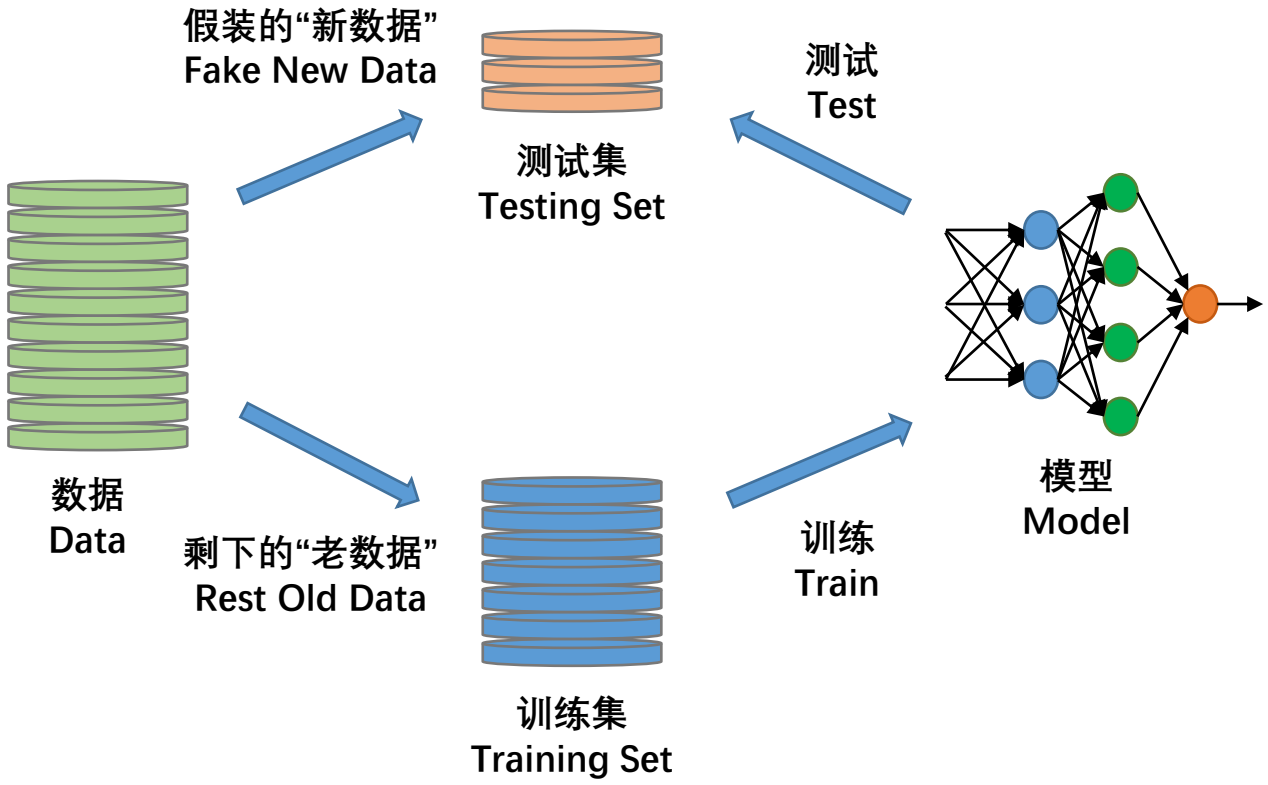
- **② 假装的新数据在逻辑上也是新数据（抽取时尽可能随机），所以在训练时一定不能用来建立模型，否则就违背了机器学习的核心逻辑，就变成了用数学技巧凑模型！**



机器学习的基本流程 Process of Machine Learning

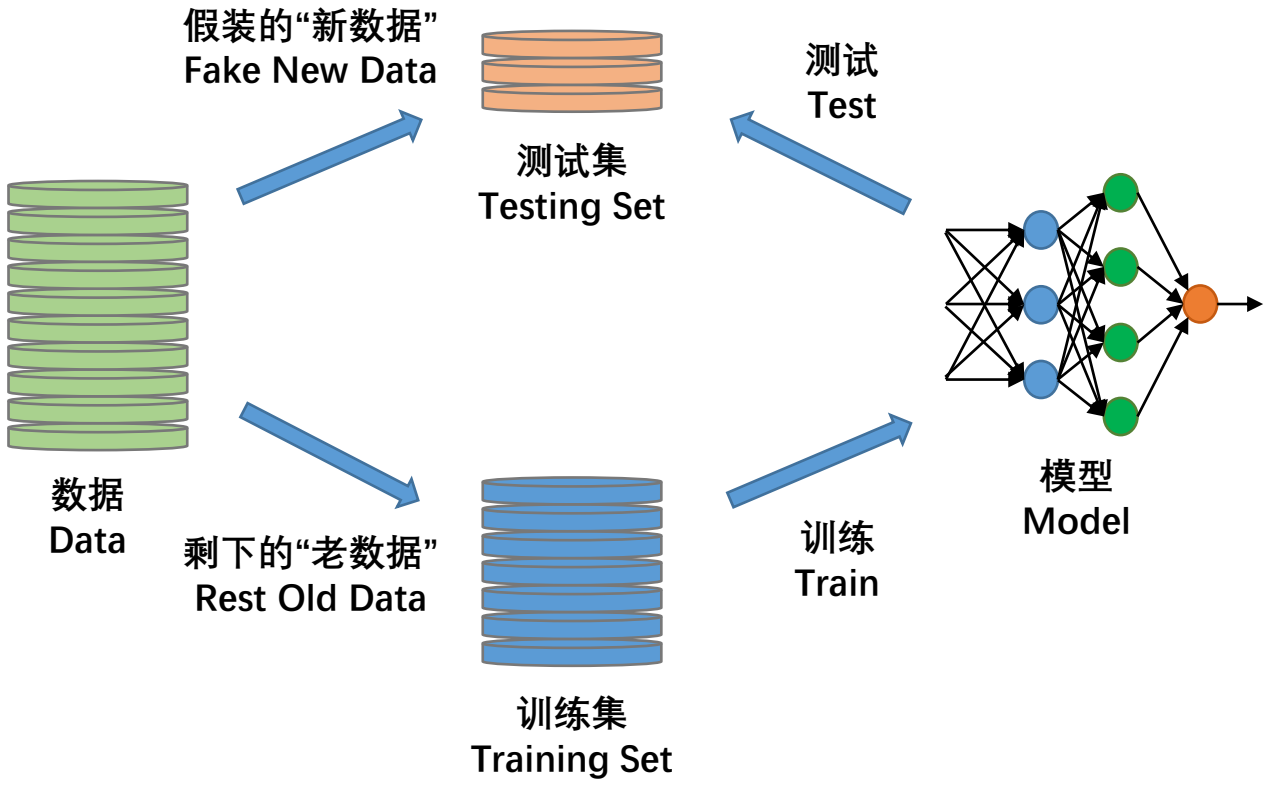
- 注意事项！Warnings!

- ③ 如果用测试集测试发现模型拟合的不好怎么办？
- ——只能重新随机抽取新数据，全部流程推倒重来。



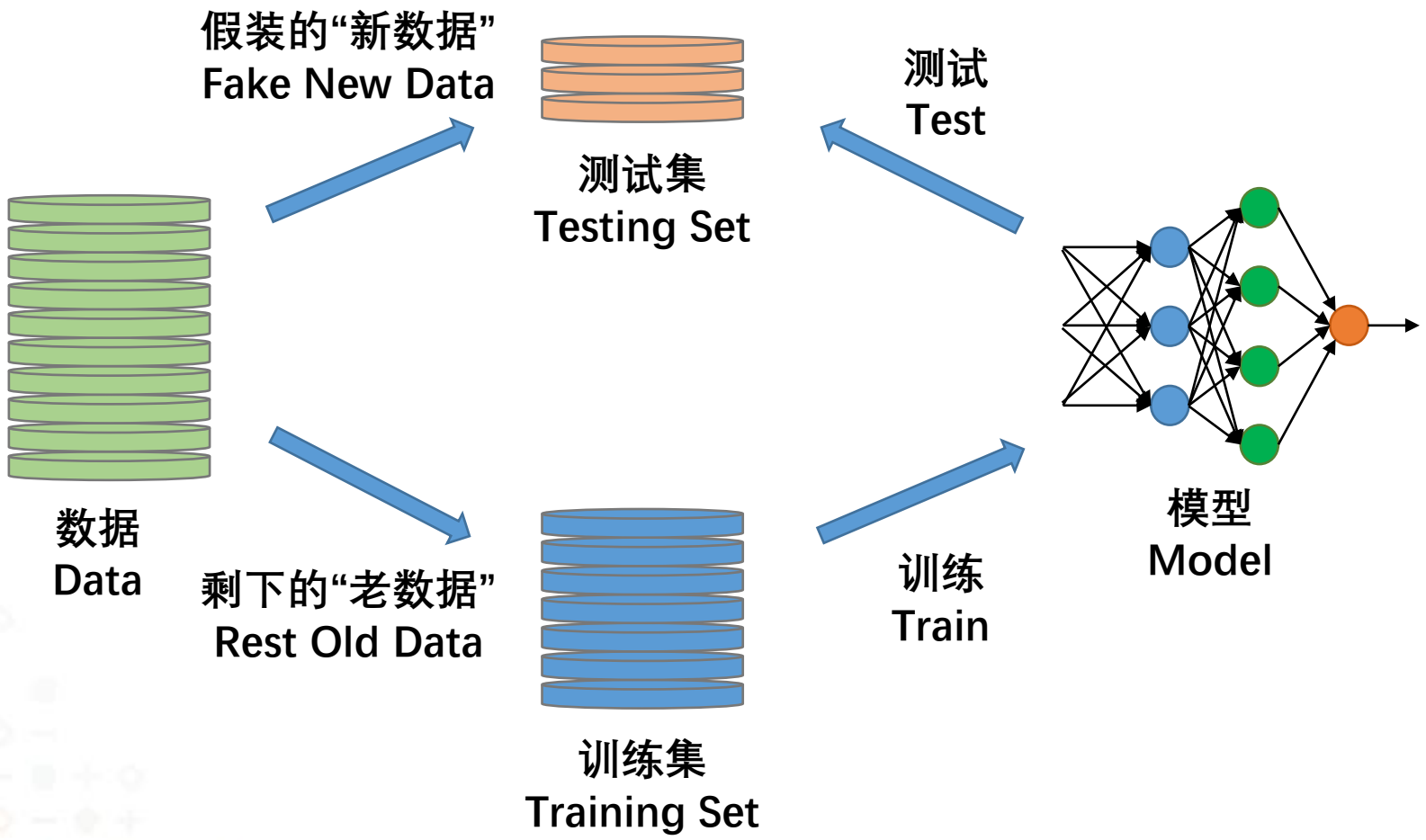
机器学习的基本流程 Process of Machine Learning

- **注意事项！Warnings!**
- ③⁺ 如果用测试集测试发现模型拟合的不好怎么办?
- **我就是想偷懒!**



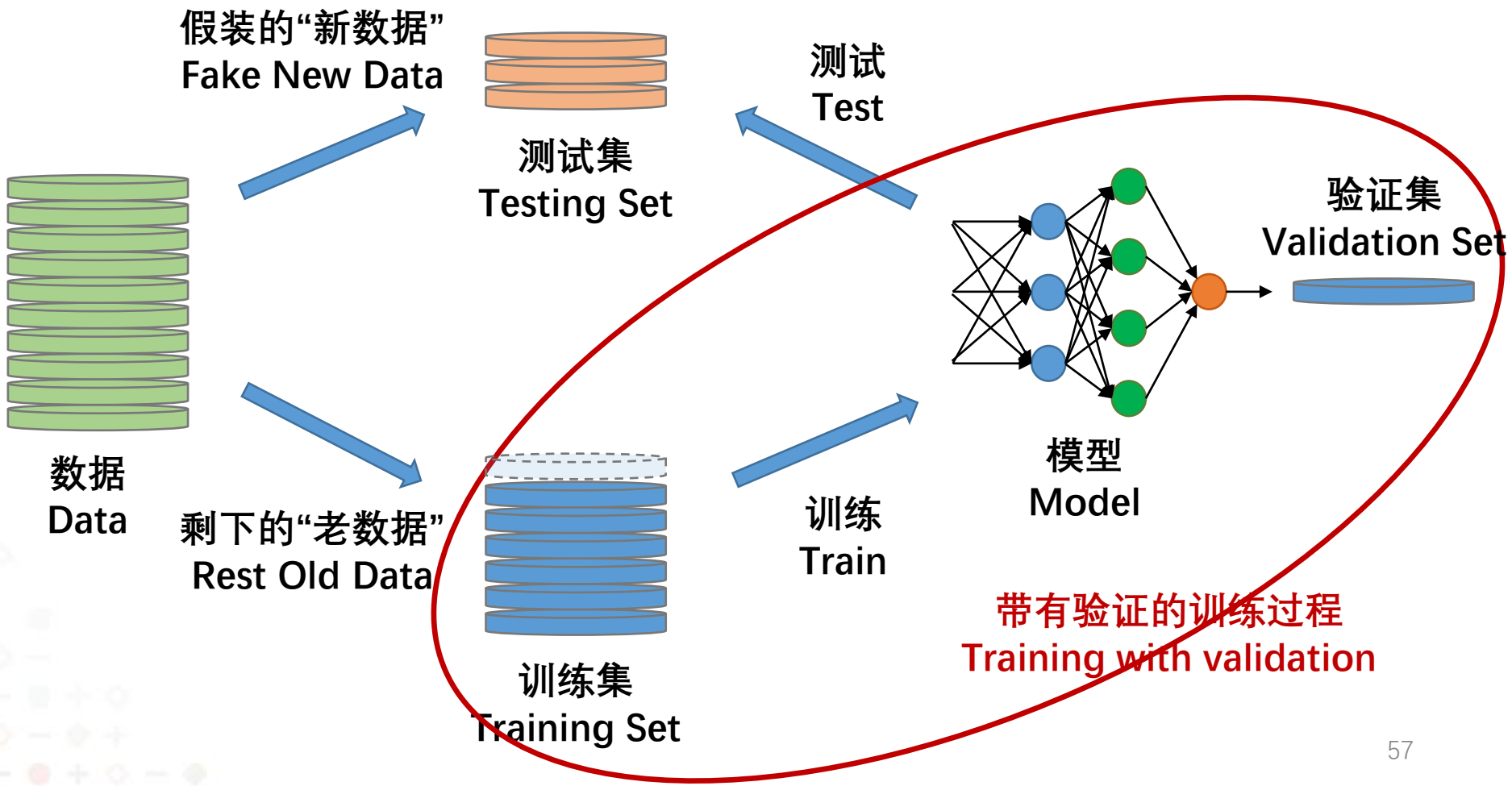
机器学习的基本流程 Process of Machine Learning

- 加入验证集（Validation Set）



机器学习的基本流程 Process of Machine Learning

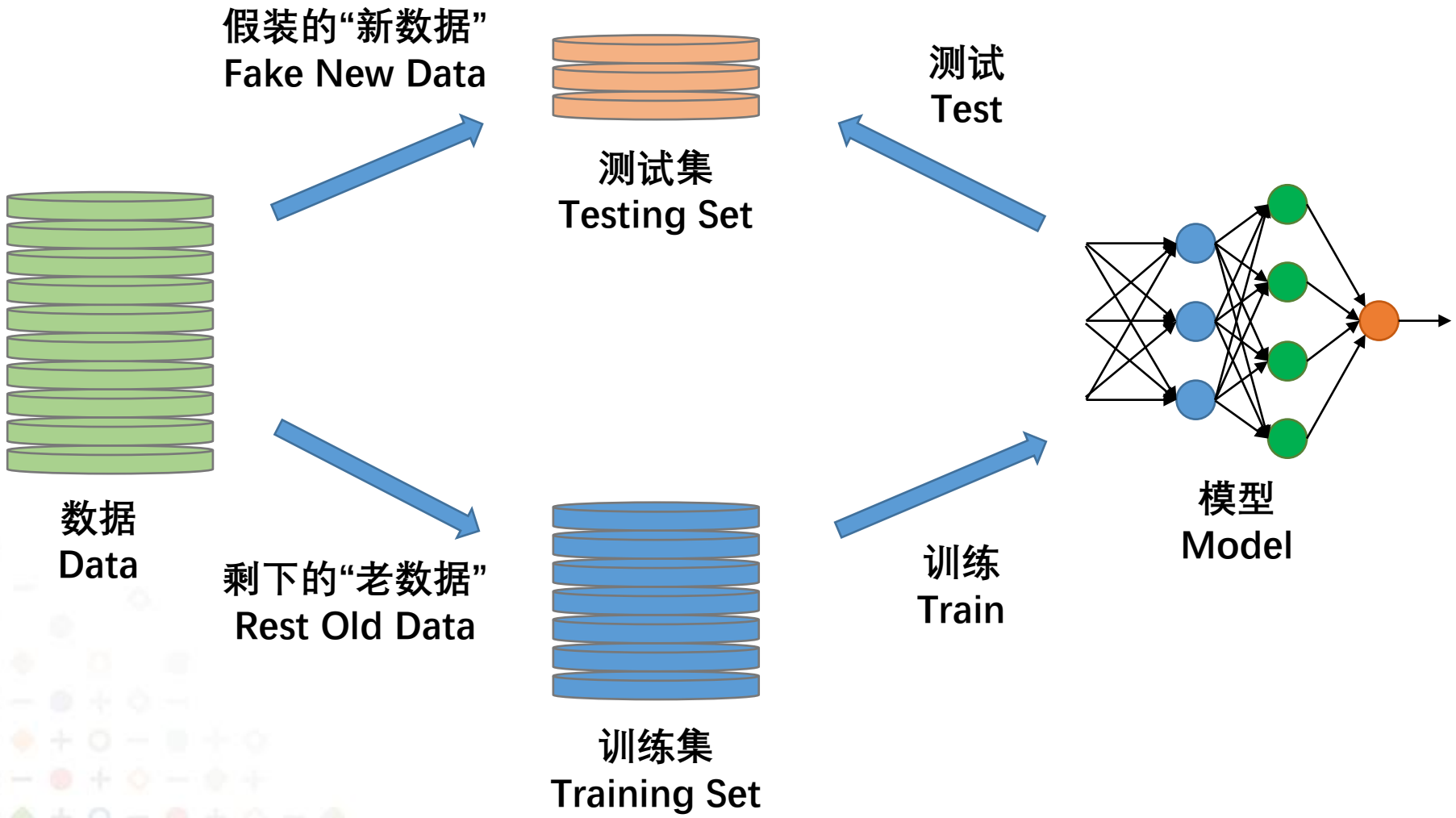
- 加入验证集（Validation Set）



机器学习的基本流程 Process of Machine Learning

- 验证集（Validation Set）
- 为了简化流程，突出核心，我们在后续的讲解中，不加入验证集

机器学习的基本流程 Process of Machine Learning



机器学习的基本流程 Process of Machine Learning

- 例
- 手写数字识别 MNIST Dataset
- 1998, Lecun Yann, Corinna Cortes, Christopher J.C. Burges

单个样本

\+ 7

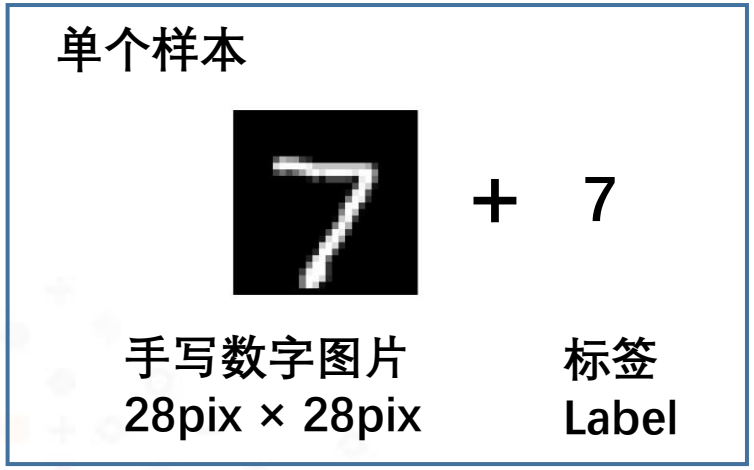
手写数字图片
28pix × 28pix

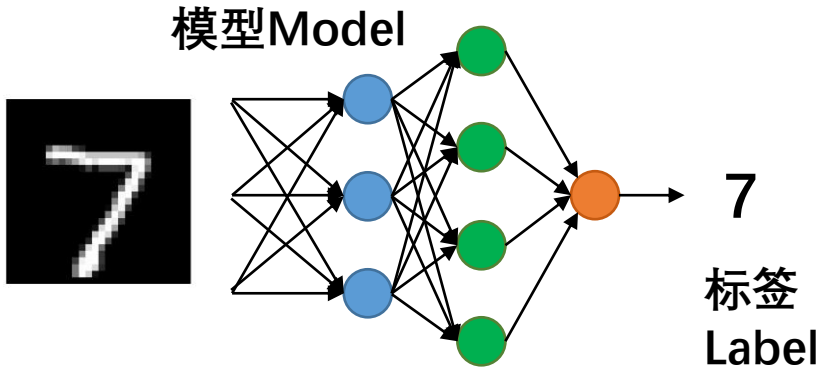
标签
Label

本页内容来源：http://yann.lecun.com/exdb/mnist/

机器学习的基本流程 Process of Machine Learning

- 例
- 手写数字识别 MNIST Dataset
- 1998, Lecun Yann, Corinna Cortes, Christopher J.C. Burges





手写数字图片

模型：输入一张手写数字图片，输出数字几

本页内容来源：http://yann.lecun.com/exdb/mnist/

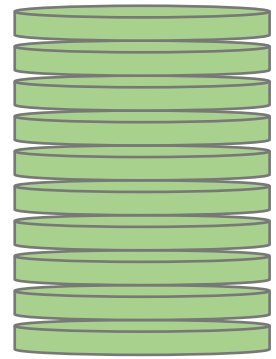
机器学习的基本流程 Process of Machine Learning

- 例
- 手写数字识别 MNIST Dataset
- 1998, Lecun Yann, Corinna Cortes, Christopher J.C. Burges

- 整个数据集 70000个样本；训练集：60000个样本；测试集10000个样本

本页内容来源：http://yann.lecun.com/exdb/mnist/

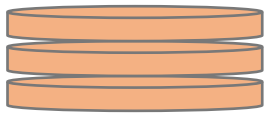
机器学习的基本流程 Process of Machine Learning



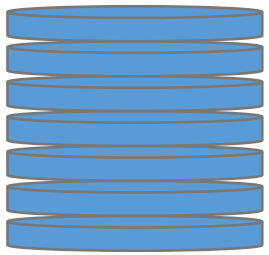
数据
Data

假装的“新数据”
Fake New Data

剩下的“老数据”
Rest Old Data



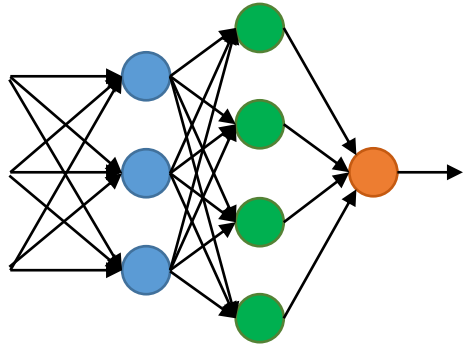
测试集 Testing Set
10000个样本



训练集 Training Set
60000个样本

测试
Test

```
Train on 60000 samples
Epoch 1/5
60000/60000 [==============================] - 2s 29us/sample - loss: 0.3037 - accuracy: 0.9111
Epoch 2/5
60000/60000 [==============================] - 1s 23us/sample - loss: 0.1464 - accuracy: 0.9575
Epoch 3/5
60000/60000 [==============================] - 1s 23us/sample - loss: 0.1122 - accuracy: 0.9663
Epoch 4/5
60000/60000 [==============================] - 1s 23us/sample - loss: 0.0902 - accuracy: 0.9731
Epoch 5/5
60000/60000 [==============================] - 1s 22us/sample - loss: 0.0775 - accuracy: 0.9757
10000/10000 [==============================] - 0s 21us/sample - loss: 0.0759 - accuracy: 0.9773
```



模型
Model

训练
Train

```
model = tf.keras.Sequential() #开始生成一个神经网络
model.add(tf.keras.layers.Flatten(input_shape = (28, 28))) #这是一个纯展平操作，把28*28的数据展评成784*1的数据
model.add(tf.keras.layers.Dense(128, activation = 'relu')) #128个神经元的全连接神经网络，激活函数是relu函数
model.add(tf.keras.layers.Dropout(0.2)) #添加一个Dropout防止过拟合
model.add(tf.keras.layers.Dense(10, activation='softmax'))
model.compile(optimizer='adam',loss='sparse_categorical_crossentropy', metrics=['accuracy']) #编译模型
```

Thanks!