机器学习与文本分析高阶

李锋
中共中央党校（国家行政学院）

机器学习

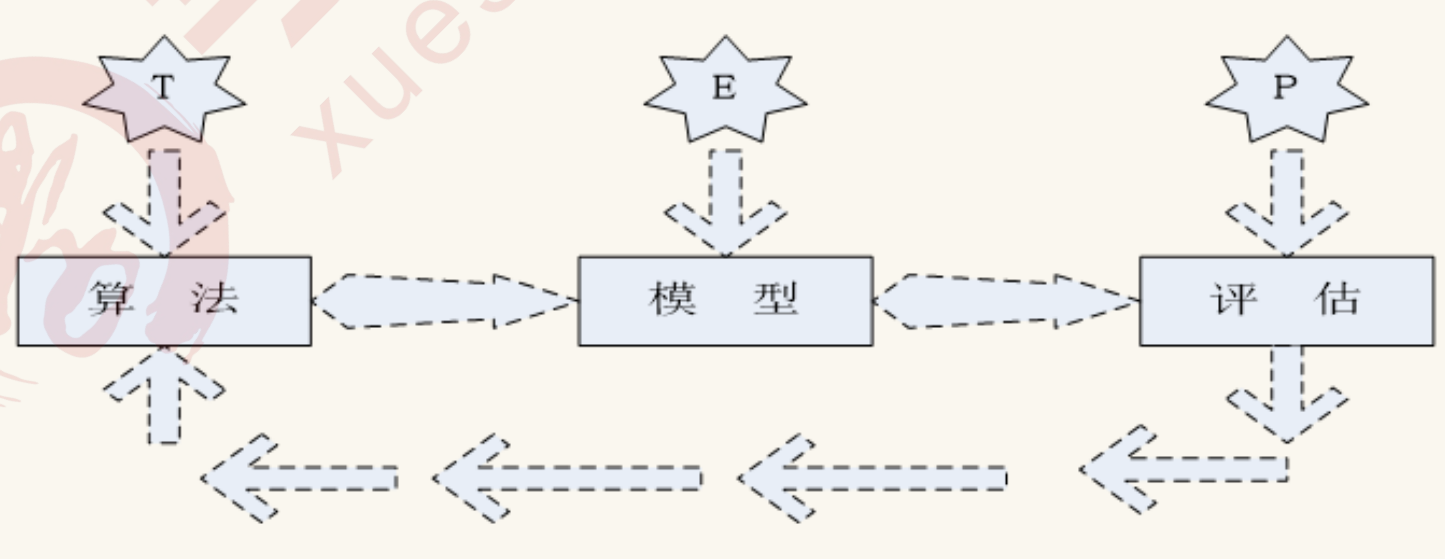
机器学习的定义

维基百科

- “机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能”。
- “机器学习是对能通过经验自动改进的计算机算法的研究”。
- “机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。”

机器学习的定义

A computer program is said to learn from experience (E) with respect to some class of tasks(T) and performance(P) measure , if its performance at tasks in T, as measured by P, improves with experience E”



机器学习与大数据

- 数据是信息和依据，背后隐含了大量不易被我们感官识别的信息、知识、规律等等。
- **机器学习的任务，就是要在基于大数据量的基础上，发掘其中蕴含并且有用的信息。**其处理的数据越多，机器学习就越能体现出优势，以前很多用机器学习解决不了或处理不好的问题，通过提供大数据得到很好解决或性能的大幅提升，如语言识别、图像设别、天气预测等等。

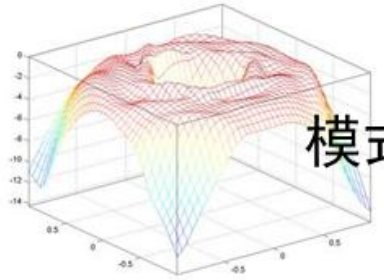
机器学习与人工智能

人工智能是计算机科学的一个分支，目的是开发一种拥有智能行为的机器，让机器像人类一样“思考”。

- Alan Turing 于 1950 年的报纸上率先提出机器人能否像人类一样思考的问题，此问题后来引出了著名的图灵测试
- 1956年的夏天，在达特茅斯，约翰·麦卡锡和克劳德·香农等共同讨论了当时计算机科学领域尚未解决的问题，第一次提出了人工智能的概念；
- 20世纪40年代，就有研究神经网络算法的文章。

机器学习是实现人工智能的方法。机器学习是目前最接近人工智能的系统。即便可以在没有机器学习的情况下创建人工智能，但这个过程将会是复杂耗时的。

机器学习在相关领域的应用



模式识别

计算机视觉

数据挖掘

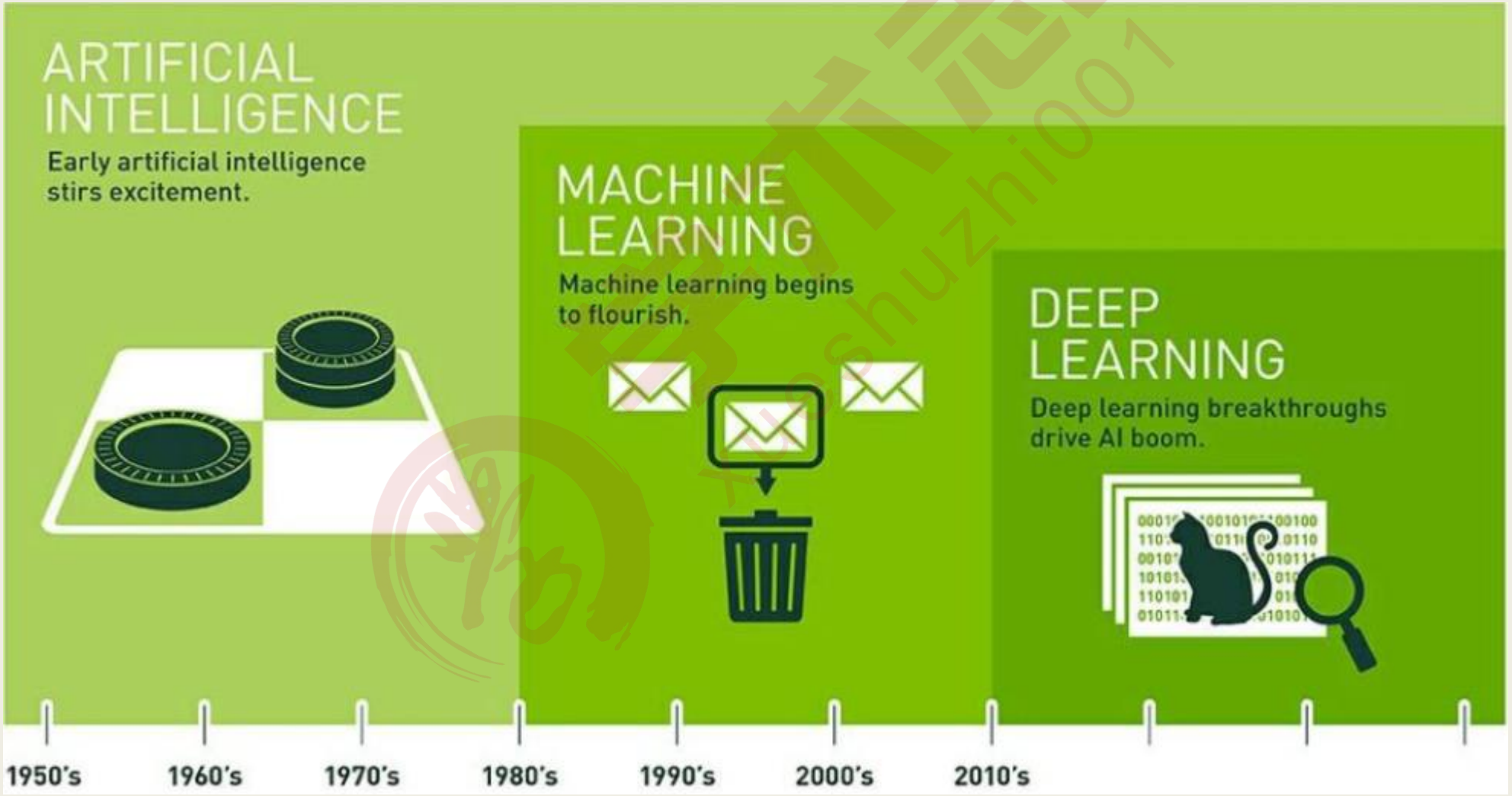
机器学习

语音识别

统计学习

自然语言处理

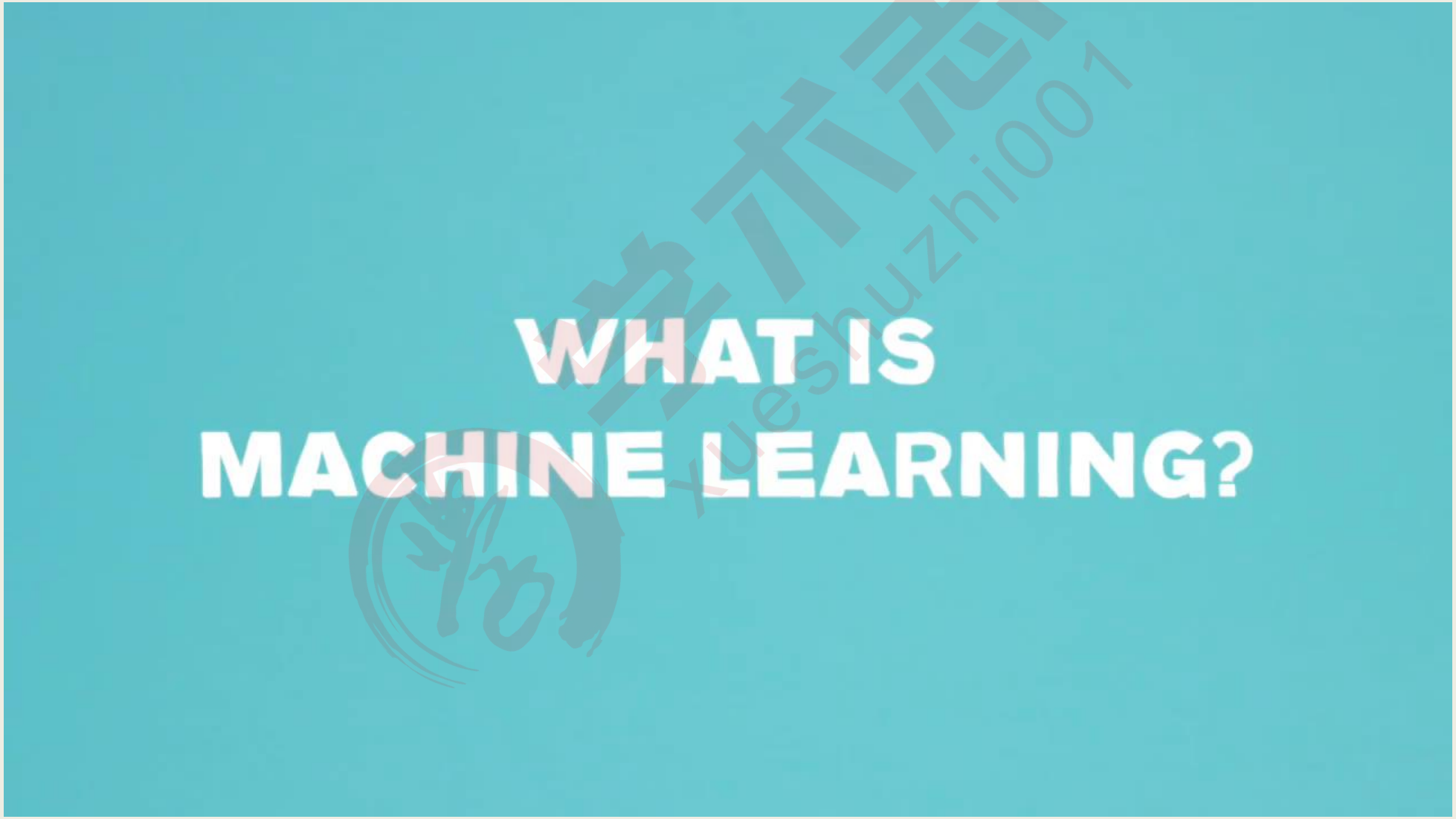
机器学习与深度学习、人工智能



机器学习与深度学习、人工智能

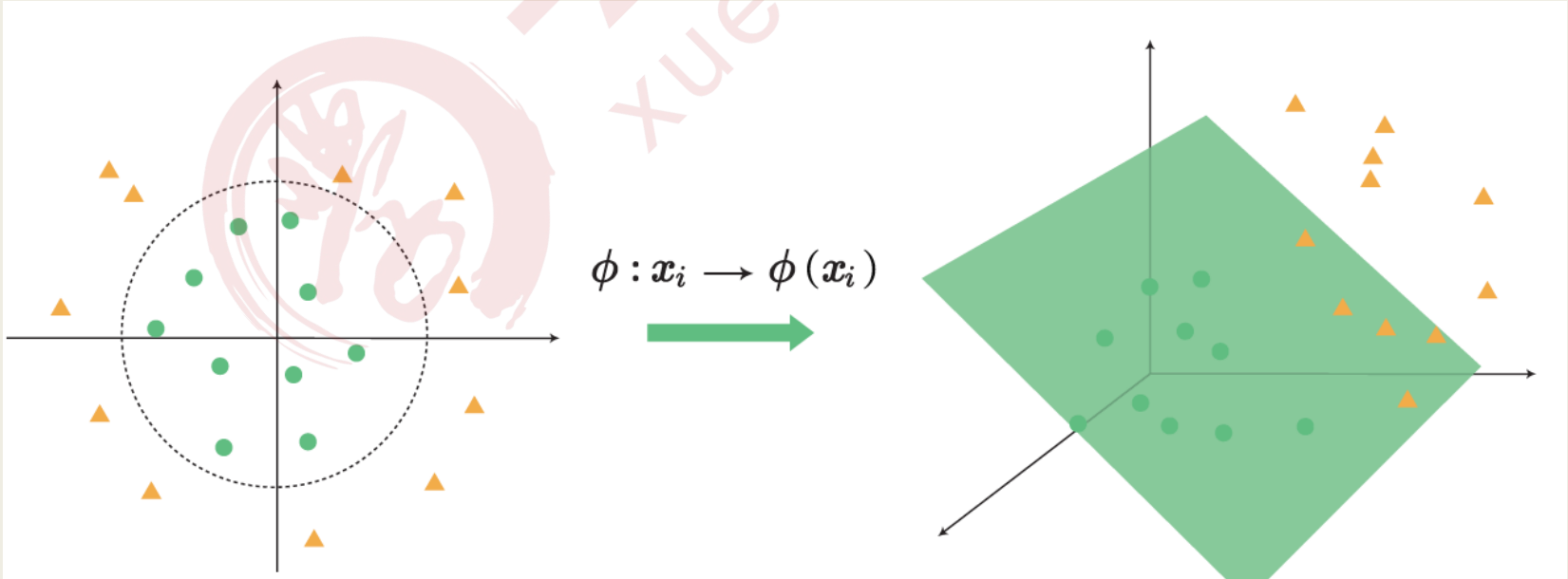
- **大数据是人工智能的基础，而使大数据转变为知识或生产力，离不开机器学习**（Machine Learning），可以说机器学习是人工智能的核心，是使机器具有类似人的智能的根本途径。
- **在很多人工智能问题上，深度学习的方法突破了传统机器学习方法的瓶颈，推动了人工智能领域的快速发展。**

机器学习



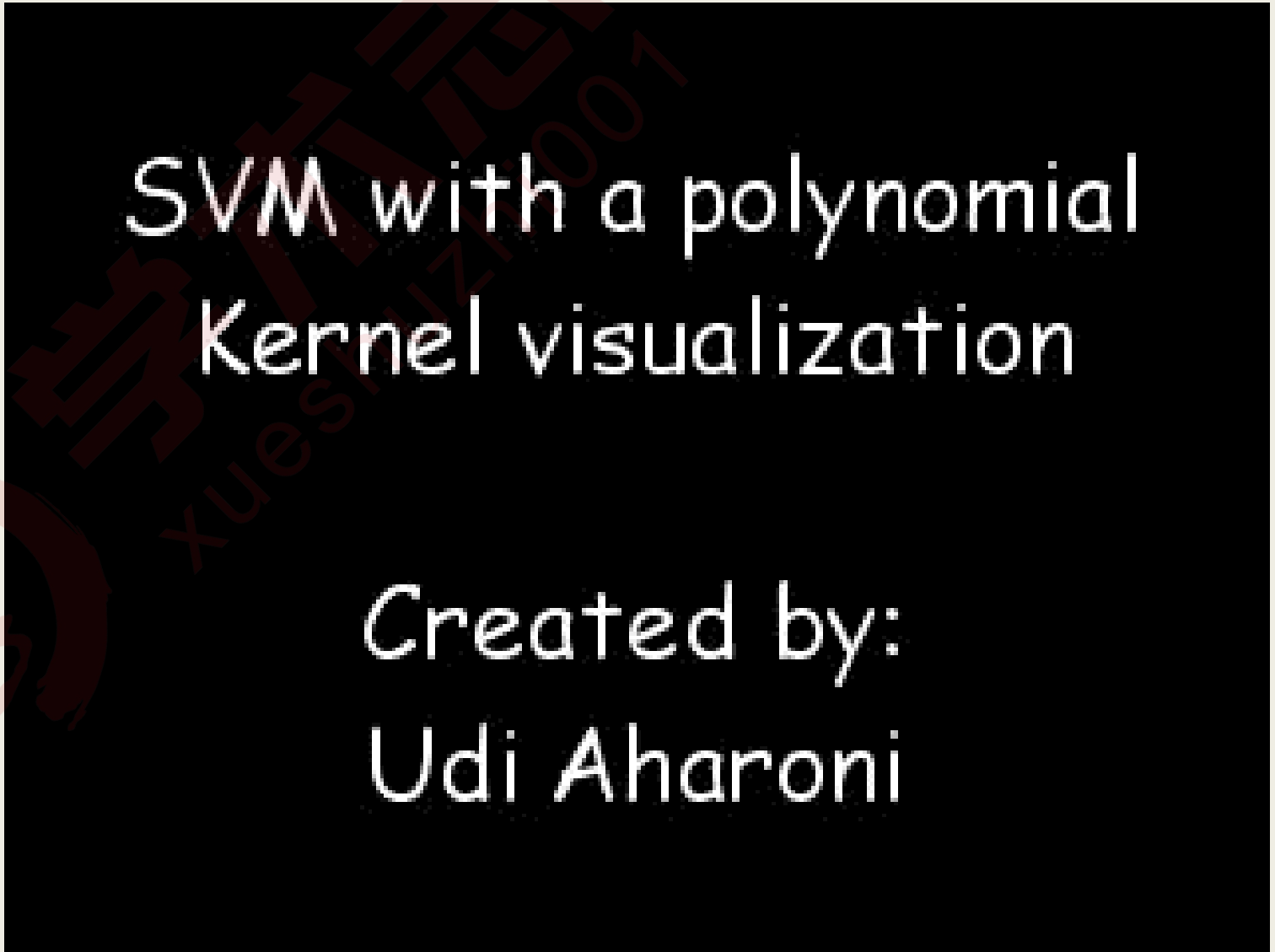
机器学习的常见方法：支持向量机

- **支持向量机算法**（Support Vector Machine，SVM)通过使用最大分类间隔来确定最优的划分超平面，以获得良好的泛化能力. SVM通过核函数的方法将低维数据映射到高维空间，并使得在高维空间中的数据是线性可分的，从而能够处理低维空间中线性不可分的情况. SVM主要应用在模式识别领域中的文本识别、文本分类、人脸识别等问题中，同时也应用到许多的工程技术和信息过滤等方面.



机器学习的常见方法：支持向量机

- **支持向量机算法**（SVM）从某种意义上来说是逻辑回归算法的强化。
- 通过与“核”的结合，支持向量机可以表达出非常复杂的分类界线，从而达成很好的的分类效果。“核”这种函数将低维空间投射到高维之上。

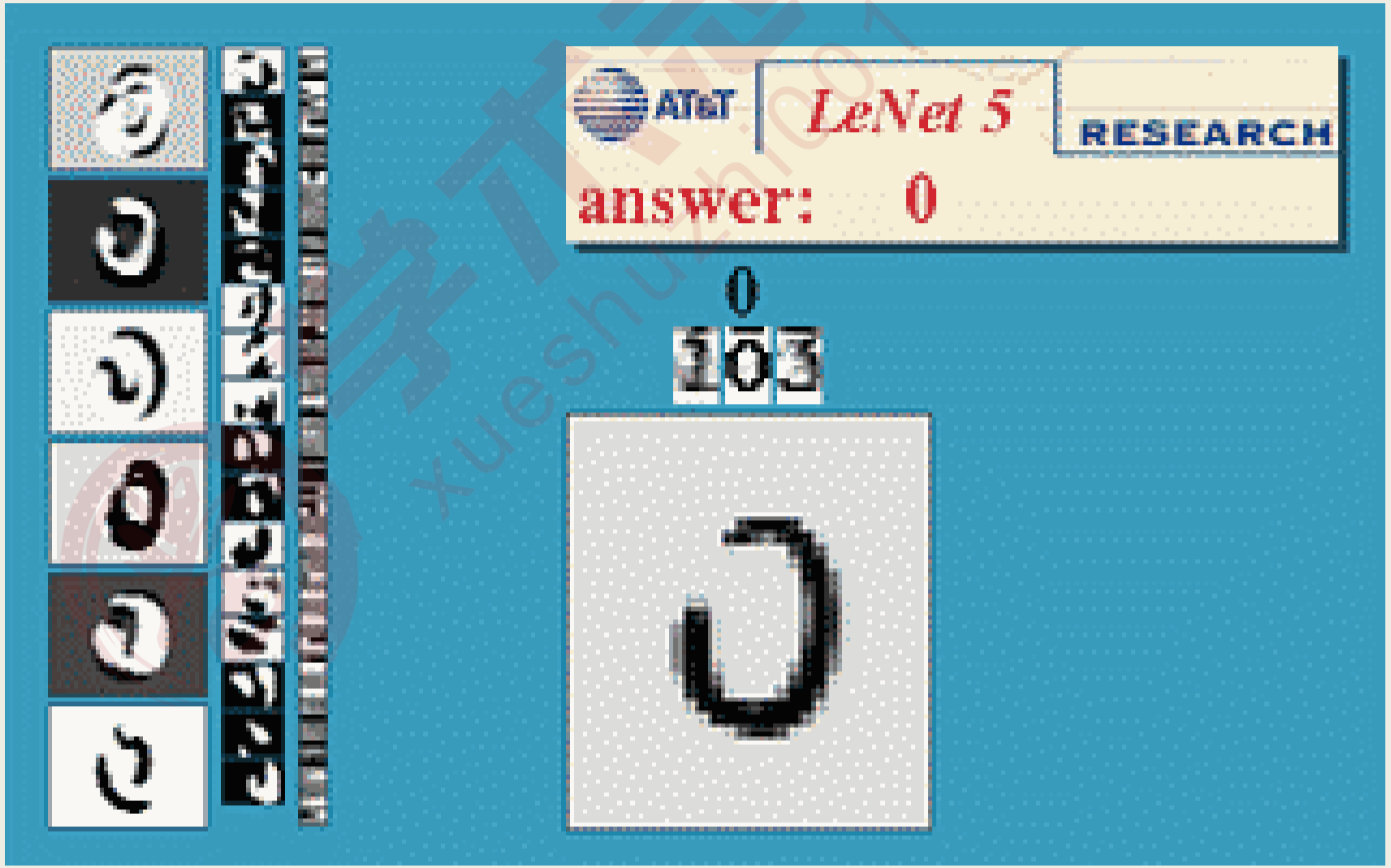


机器学习的常见方法

- **神经网络**(也称之为人工神经网络，ANN)算法，在80年代流行，90年代衰落，在近期携“深度学习”之势，再次被人重视。

机器学习的常见方法

- **神经网络**在图像识别领域的著名应用：Lanet(一个基于多个隐层构建的神经网络)

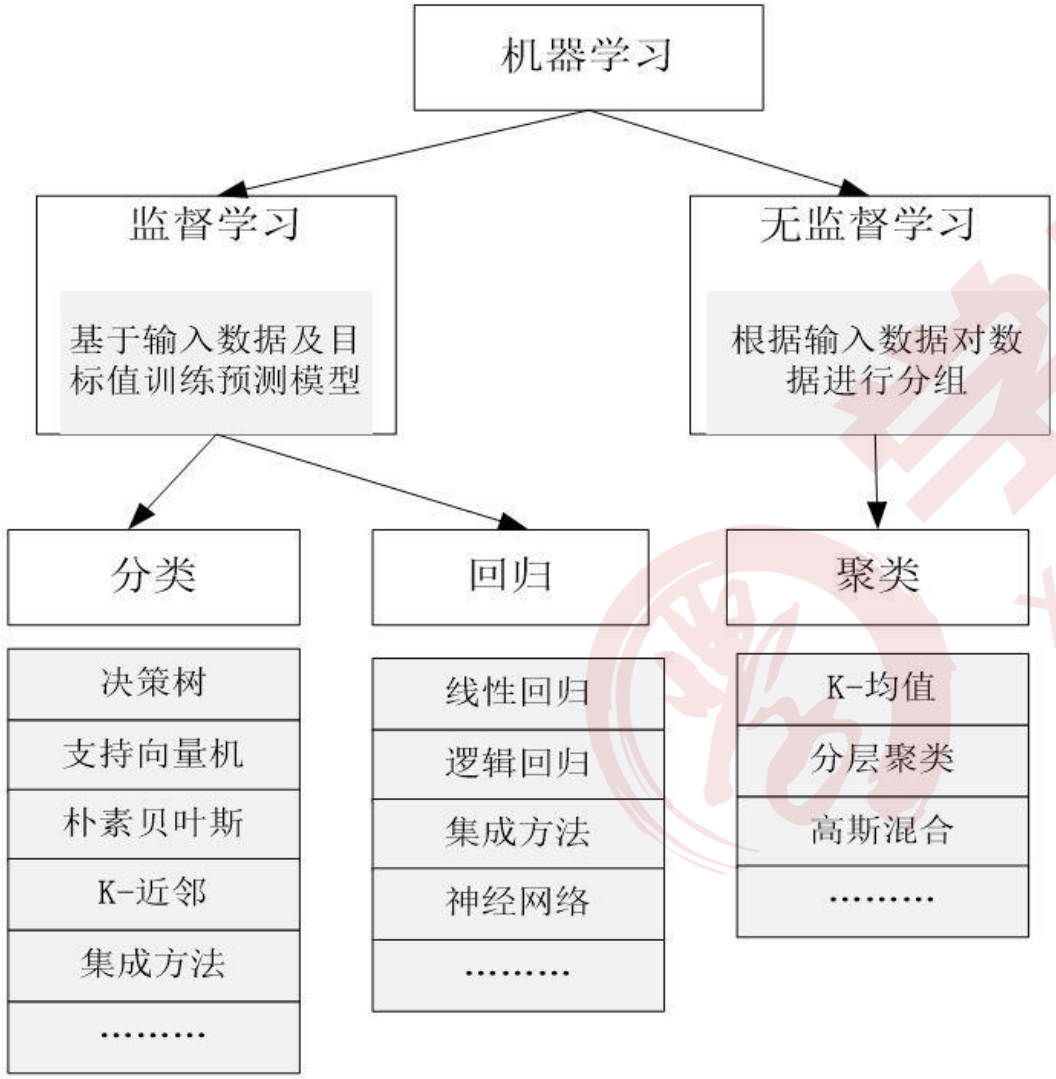


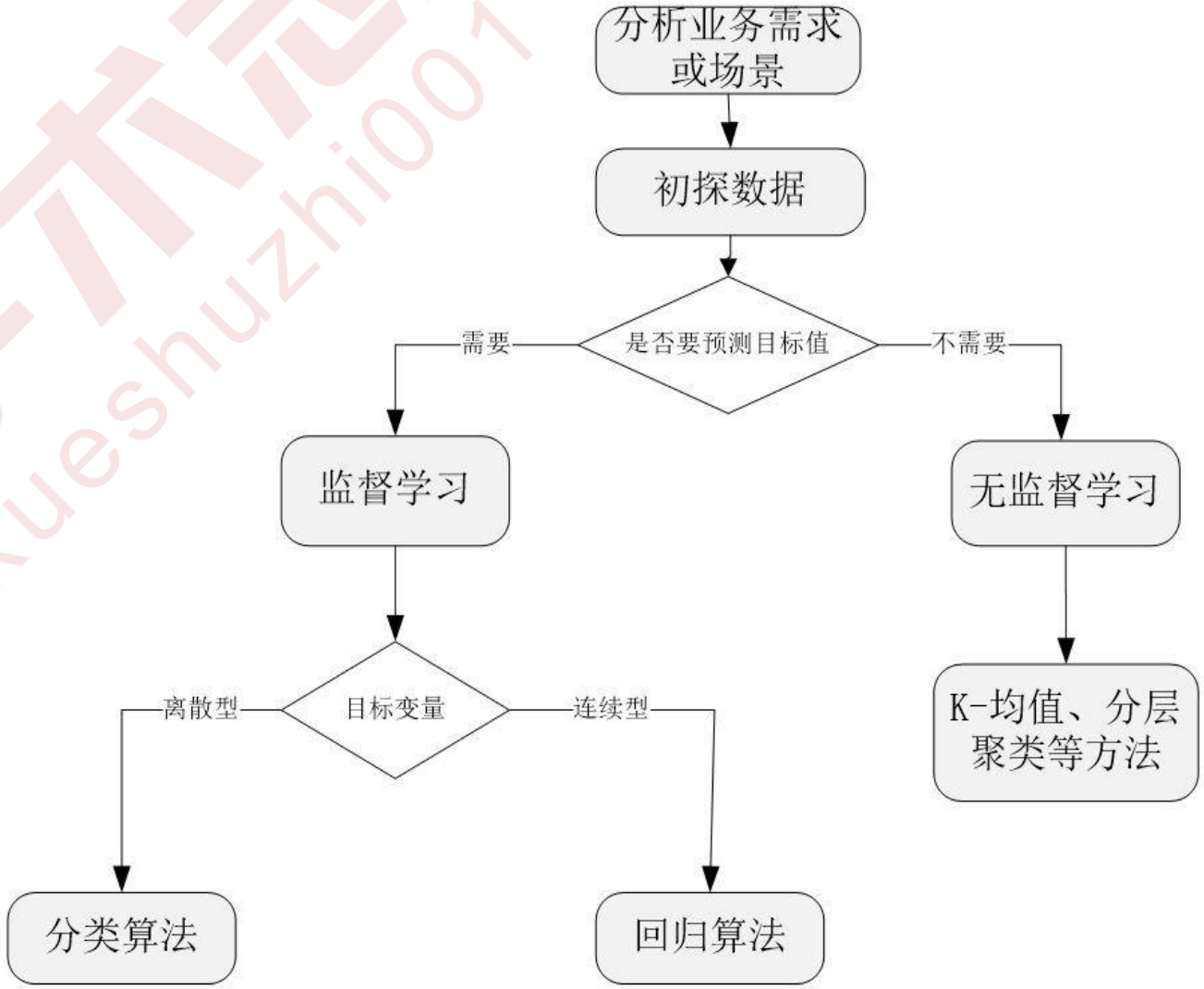
机器学习的任务

机器学习基于数据，以此获取新知识、新技能：

- 分类就是将新数据划分到合适的类别中，一般用于类别型的目标特征，如果目标特征为连续型，则往往采用回归方法。这两种方法都是先根据标签值或目标值建立模型或规则，然后利用这些带有目标值的数据形成的模型或规则，对新数据进行识别或预测。这两种方法都属于监督学习。
- 把相似或相近的数据划分到相同的组里，聚类就是解决这一类问题的方法之一。

机器学习的任务

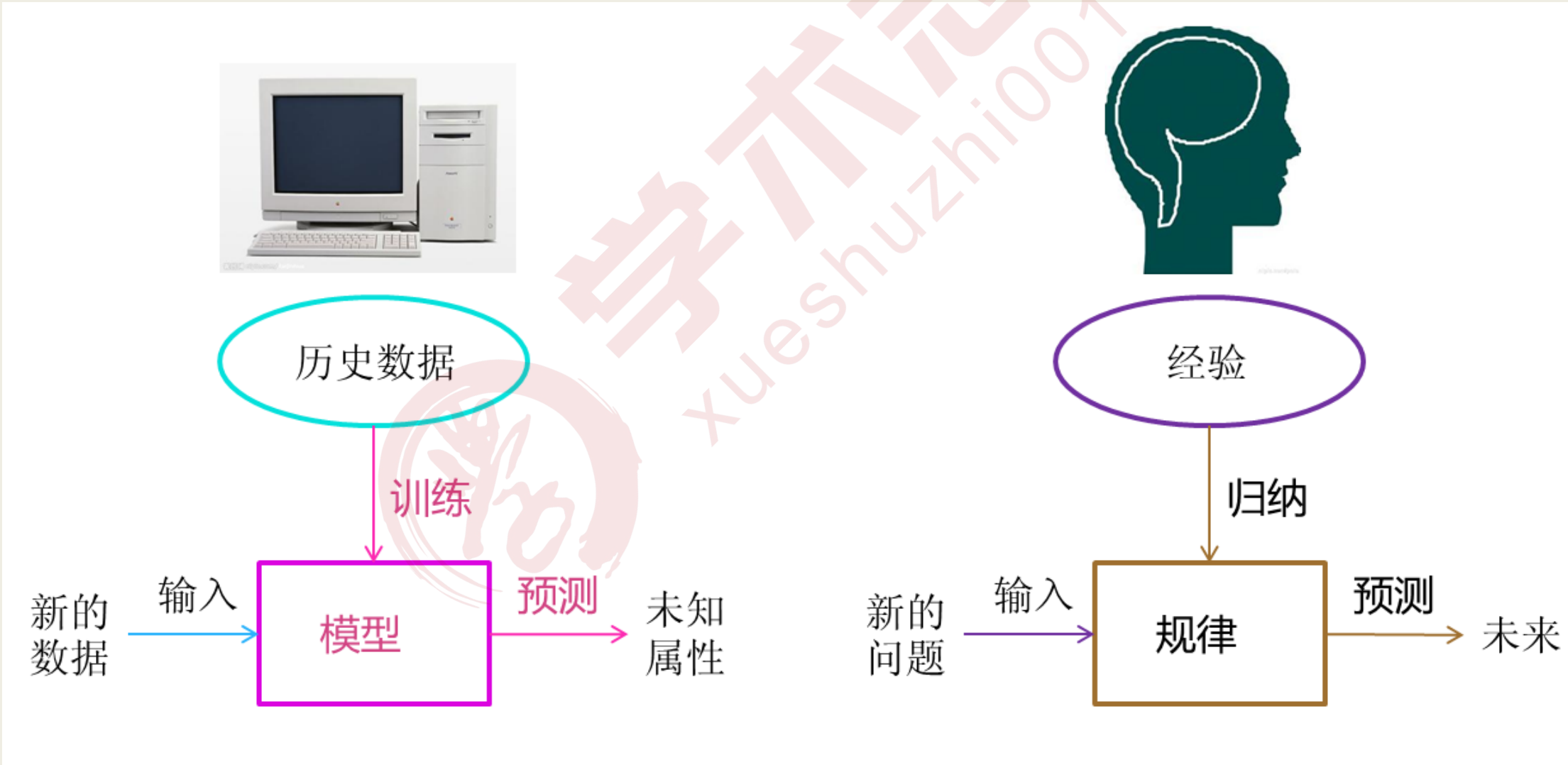

机器学习
监督学习
基于输入数据及目标值训练预测模型
无监督学习
根据输入数据对数据进行分组
分类
回归
聚类
决策树
支持向量机
朴素贝叶斯
K-近邻
集成方法
……
线性回归
逻辑回归
集成方法
神经网络
……
K-均值
分层聚类
高斯混合
……


分析业务需求或场景
初探数据
是否要预测目标值
需要
不需要
监督学习
无监督学习
目标变量
离散型
连续型
K-均值、分层聚类等方法
分类算法
回归算法

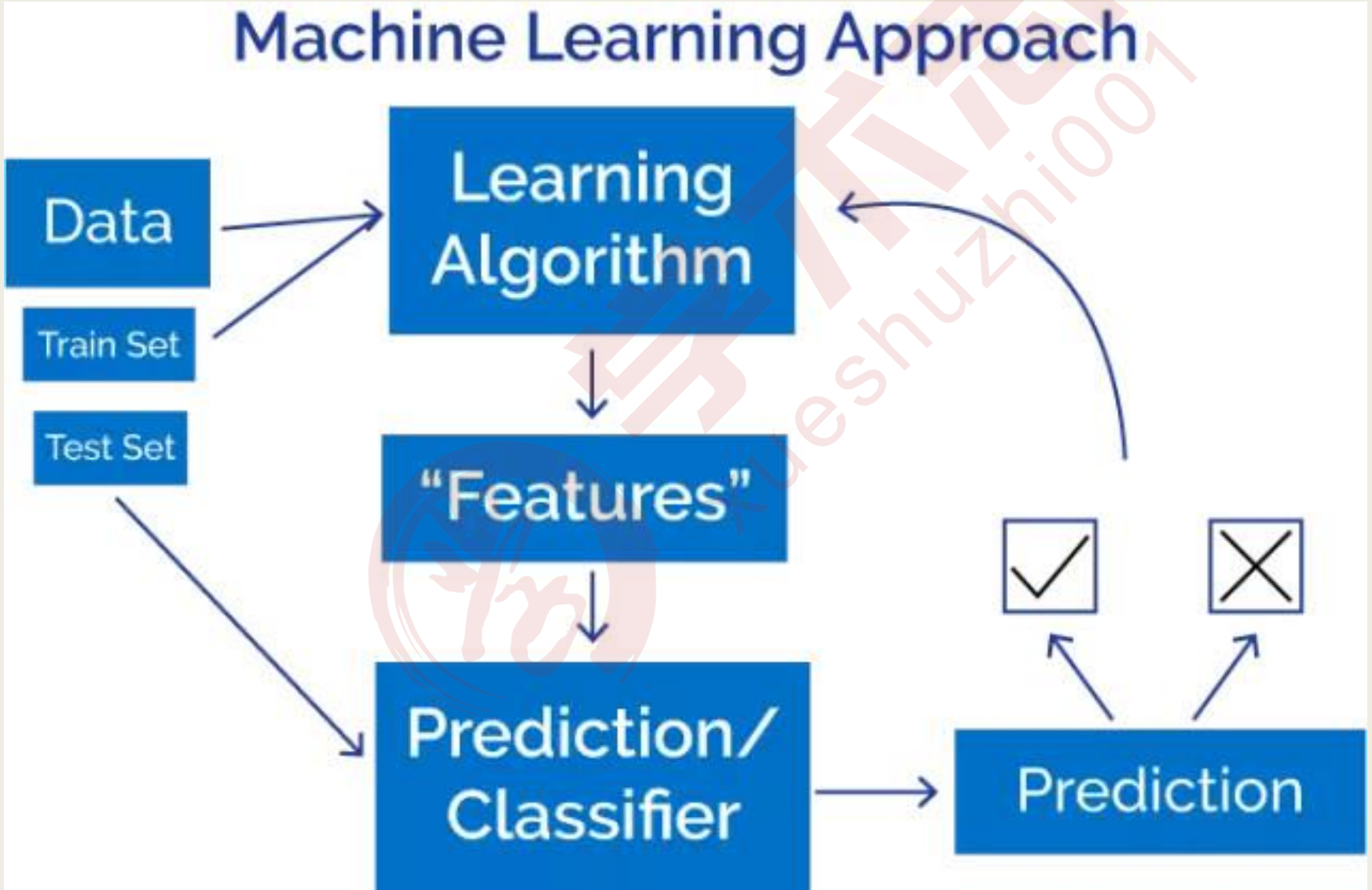
机器学习的常见方法

- 回归算法（线性回归；逻辑回归）
- K近邻（KNN）
- 支持向量机
- 神经网络
- 随机森林
- 朴素贝叶斯
- 聚类算法
- 降维算法
- 推荐算法

机器学习的过程



机器学习的流程


Machine Learning Approach
Data
Train Set
Test Set
Learning Algorithm
"Features"
Prediction/ Classifier
Prediction

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from $X \in \mathcal{X}$
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains unknown quantity with a prescribed frequency
directed acyclic graph	Bayes net	multivariate distribution with specified conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update subjective beliefs
frequentist inference	—	statistical methods for producing point estimates and confidence intervals with guarantees on frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

机器学习的实践-Python的sklearn

自2007年发布以来，scikit-learn已经成为Python重要的机器学习库了。scikit-learn简称sklearn，支持包括分类、回归、降维和聚类等机器学习算法，还包含了特征提取、数据处理和模型评估等模块。

安装方法：conda install scikit-learn

机器学习的实践-Python的sklearn

- Sklearn是一个机器学习的python库，里面包含了几乎所有常见的机器学习与数据挖掘的各种算法。

- 具体的，它常见的包括数据预处理（preprocessing）（正则化，归一化等），特征提取（feature_extraction）（TFIDF等），降维（decomposition）（PCA等），以及常见的机器学习算法（分类、聚类、回归），更特别的，它也包括了评估（混淆矩阵与PRF及Acc值）和参数优化等（GridSearchCV），甚至是交叉验证（cross_validation）等都包含在内，可谓是机器学习整个流程都有了。

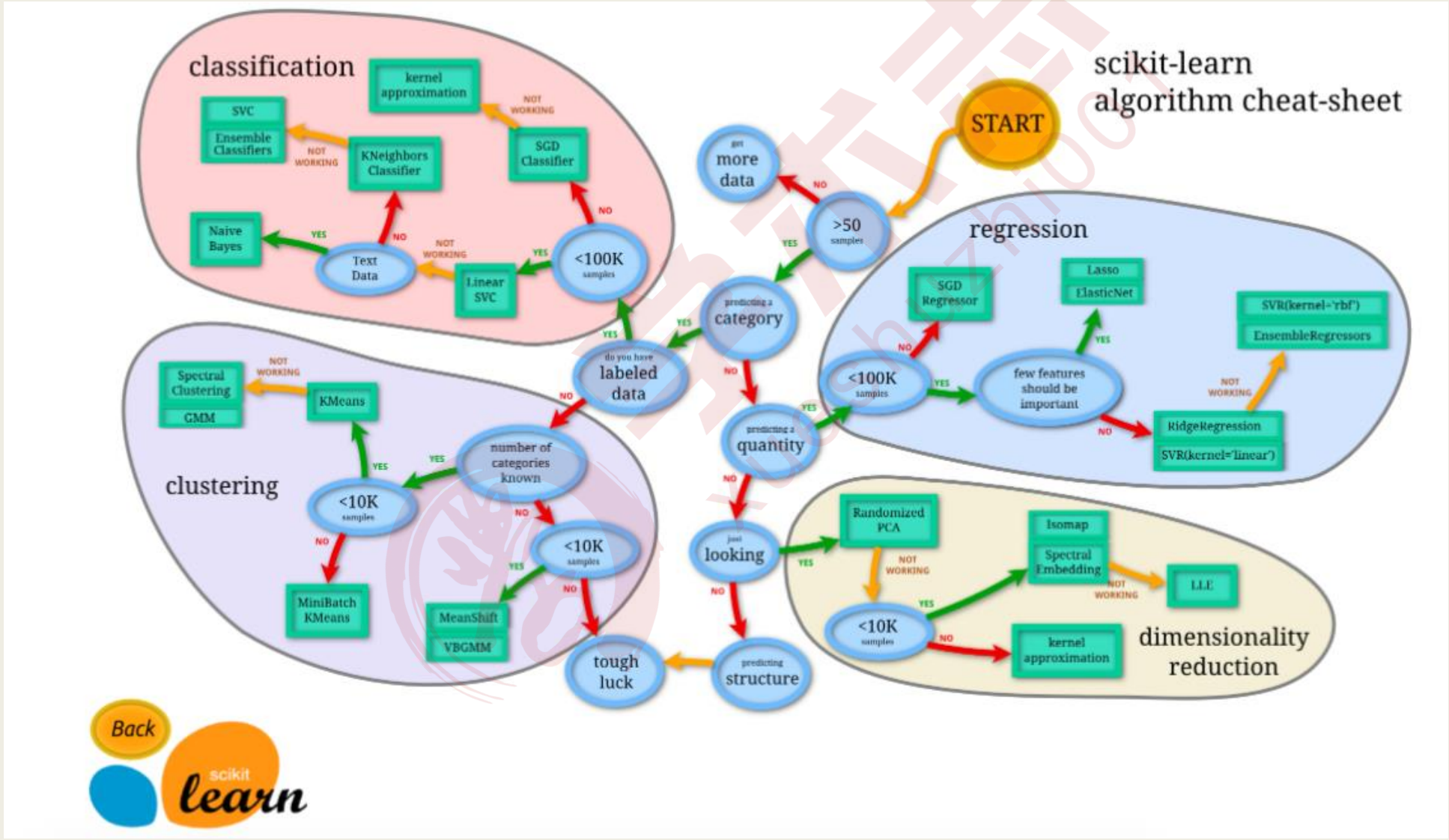
机器学习：模型选择

1）选择与业务目标一致的模型；

2）选择与训练数据和特征相符的模型。

- 训练数据少，High Level特征多，则使用“复杂”的非线性模型（流行的GBDT、Random Forest等）；
- 训练数据很大量，Low Level特征多，则使用“简单”的线性模型（流行的LR、Linear-SVM等）。

机器学习的实践-Python的sklearn



机器学习：数据预处理环节

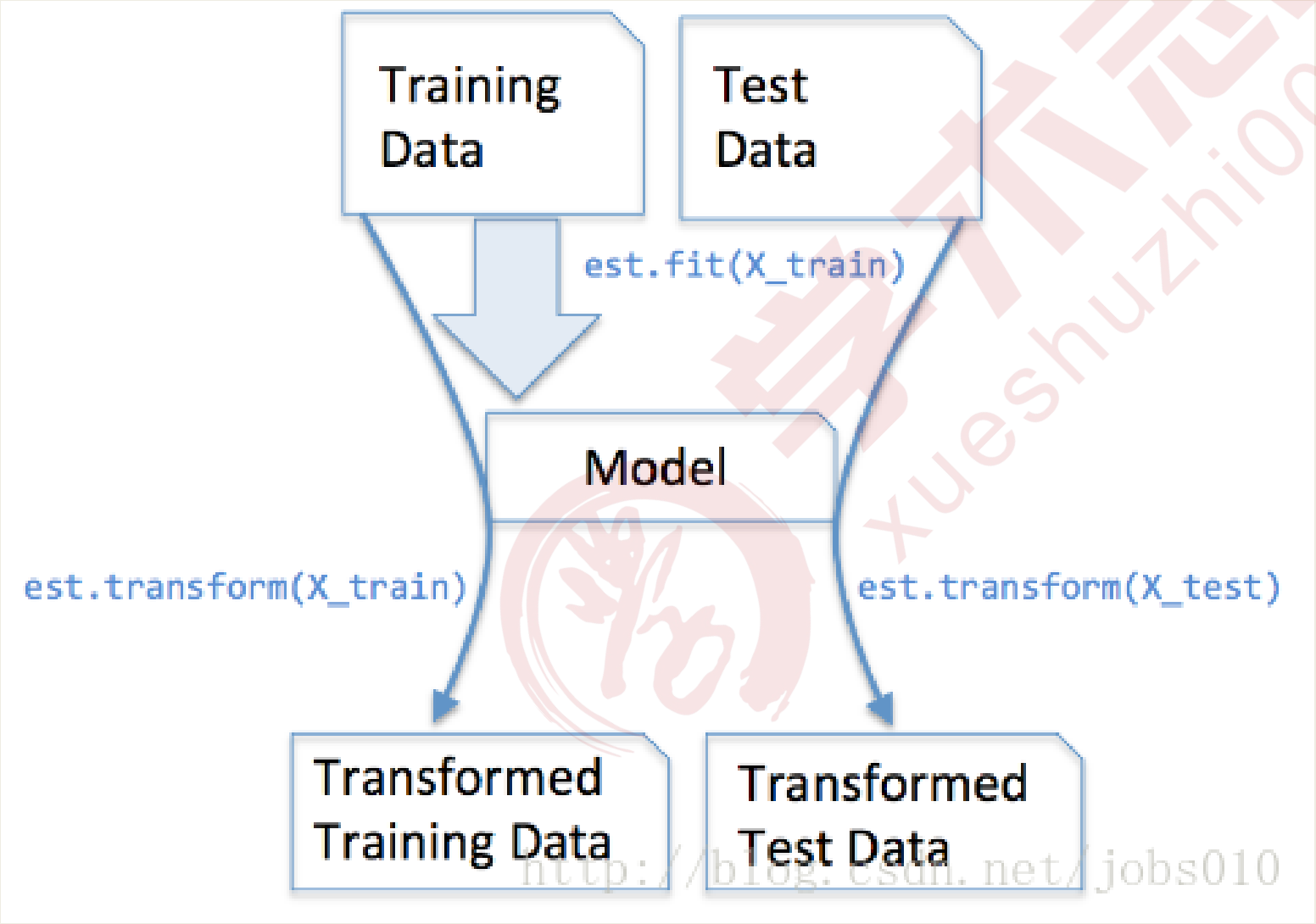
- 待解决问题的数据本身的分布尽量一致；
- 训练集/测试集分布与线上预测环境的数据分布尽可能一致，这里的分布是指（x,y）的分布，不仅仅是y的分布；
- y数据噪音尽可能小，尽量剔除y有噪音的数据；

sklearn.preprocessing包提供了几个常见的实用函数和变换器类，以将原始特征向量转换为更适合分类器或者回归器的表示，具体包括**标准化，正则化等、归一化**

机器学习：特征抽取

完成数据筛选和清洗后，就需要对数据抽取特征，就是完成输入空间到特征空间的转换（见下图）。针对线性模型或非线性模型需要进行不同特征抽取，线性模型需要更多特征抽取工作和技巧，而非线性模型对特征抽取要求相对较低。

Sklearn的特征处理环节



以CountVectorizer为例，词频矩阵

（1）fit(raw_documents) :拟合原始数据，生成文档中有价值的词汇表；

（2）transform(raw_documents):使用符合fit的词汇表或提供给构造函数的词汇表，从原始文本文档中提取词频，转换成词频矩阵。。

（3）fit_transform(raw_documents, y=None):学习vocabulary dictionary并返回term - document矩阵(稀疏矩阵)。

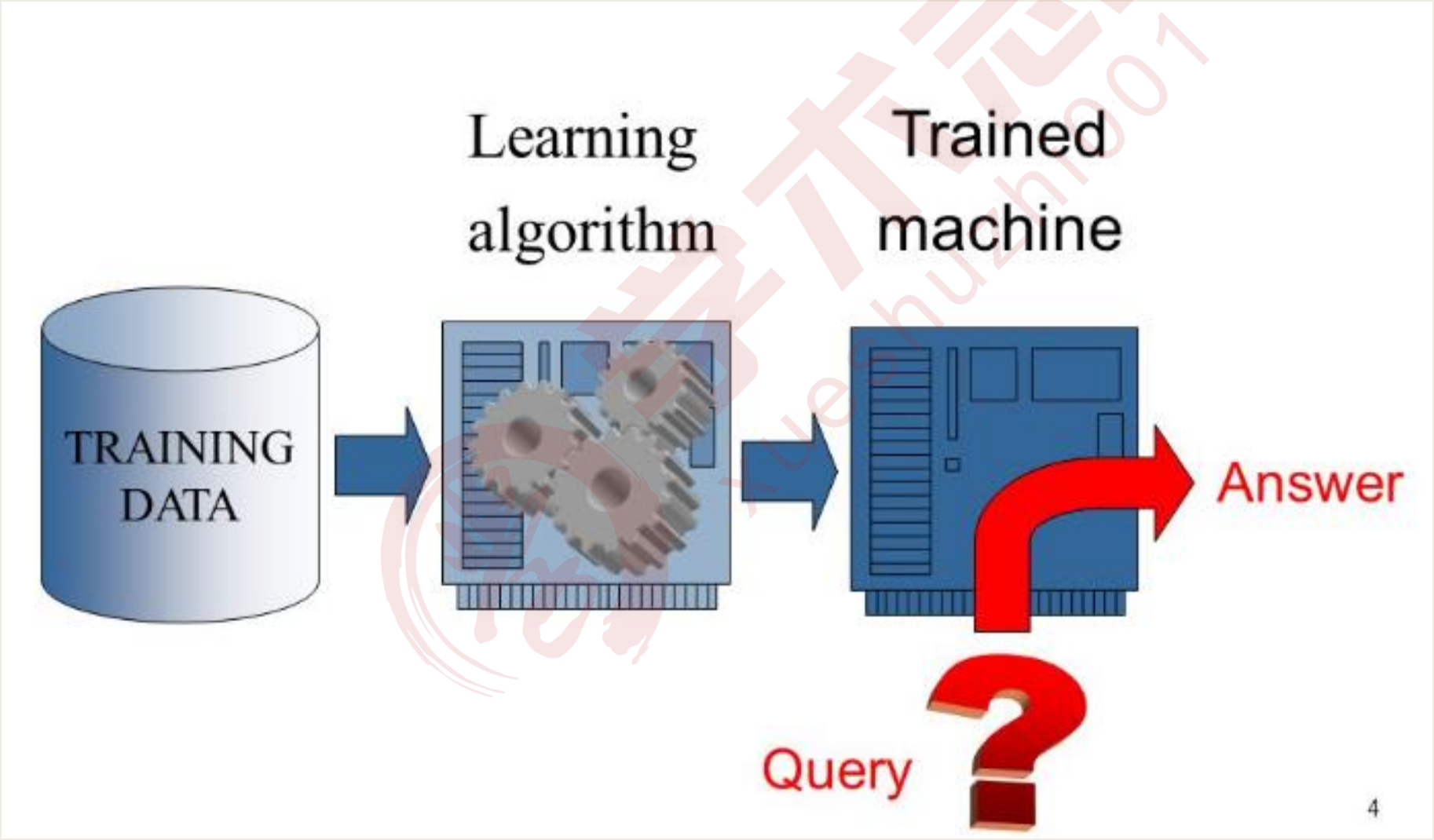
Python应用中的注意事项

- 只有有监督的转换类的fit和transform方法才需要特征和目标值两个参数，即有监督学习的算法fit(x,y)传两个参数。
- 无监督学习的算法是fit(x)，即传一个参数，比如降维、特征提取、标准化

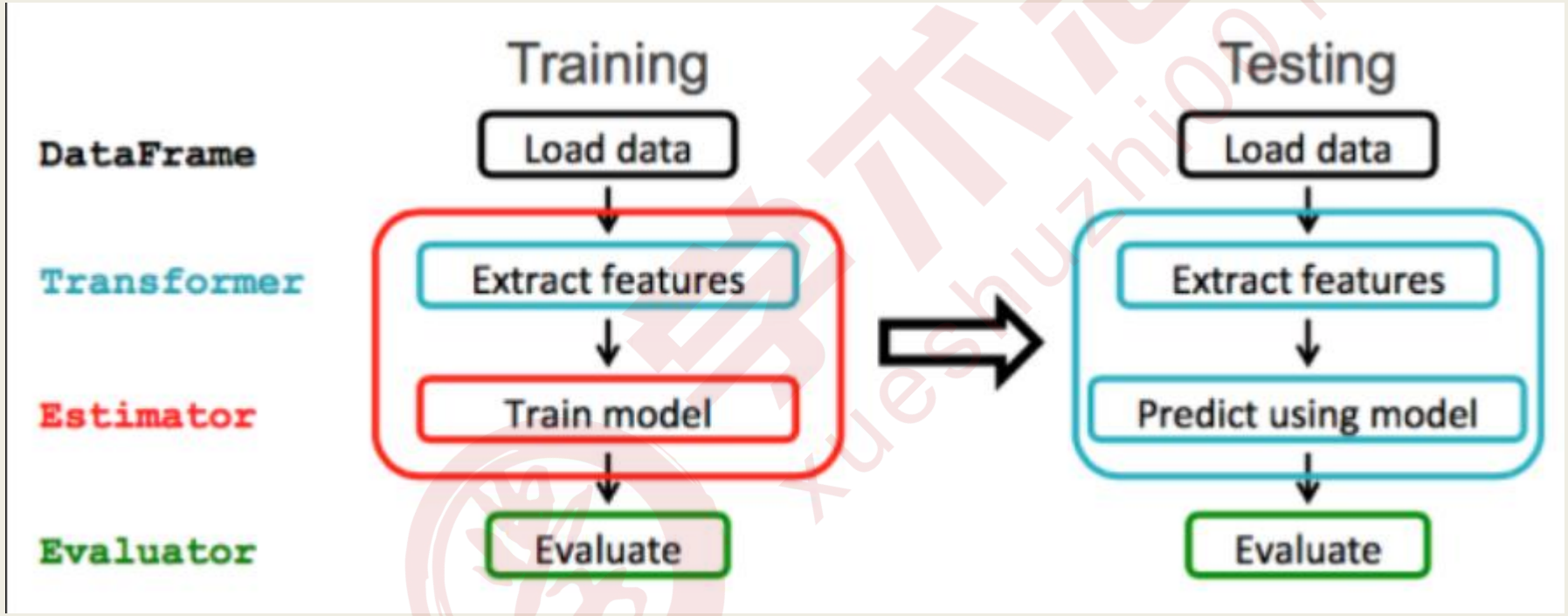
有监督学习中注意：

- 必须先用fit_transform(trainData)，之后再transform(testData)
- 如果直接transform(testData)，程序会报错

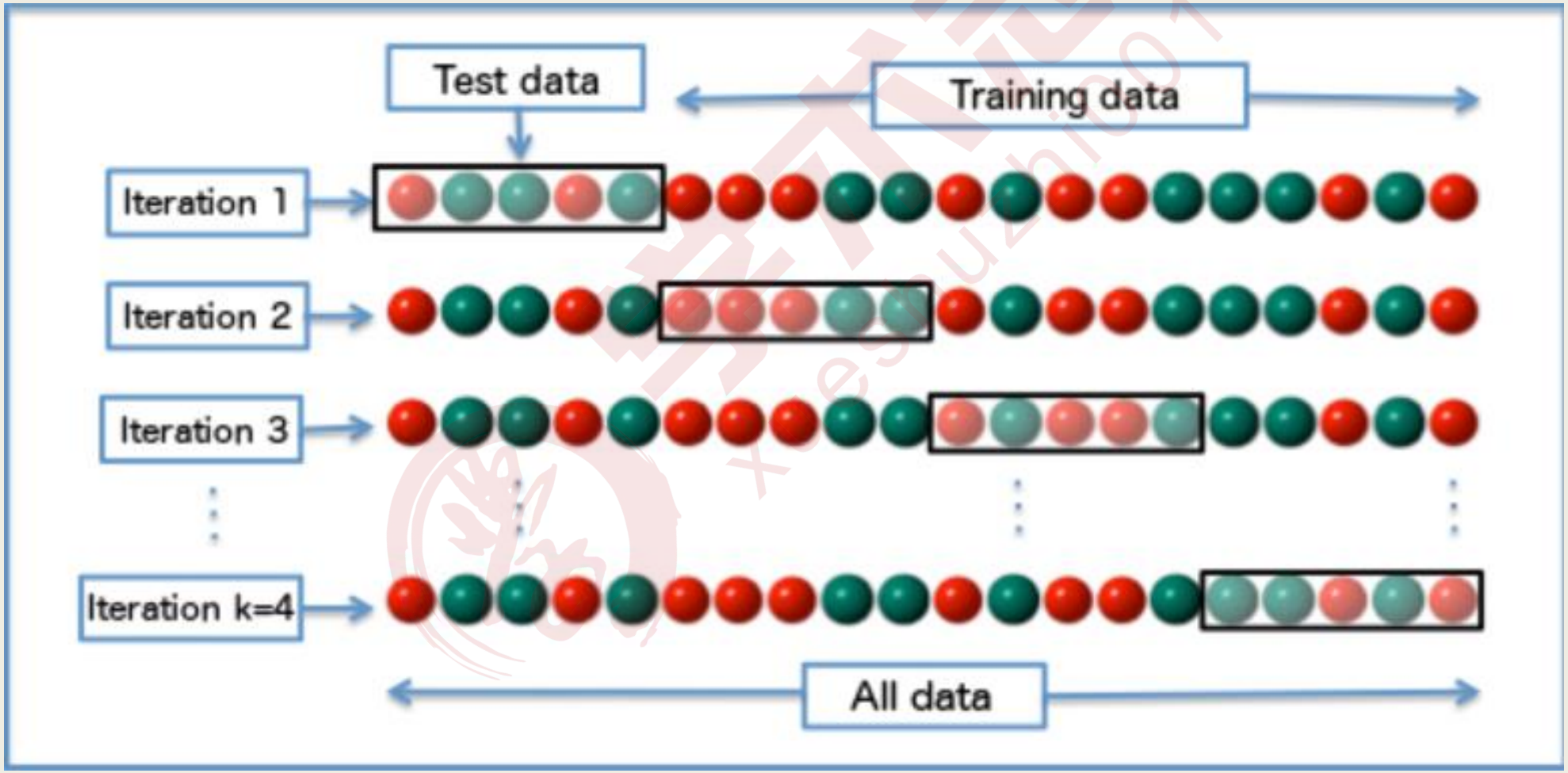
机器学习：训练模型


Learning
algorithm
Trained
machine
TRAINING
DATA
Answer
Query
4

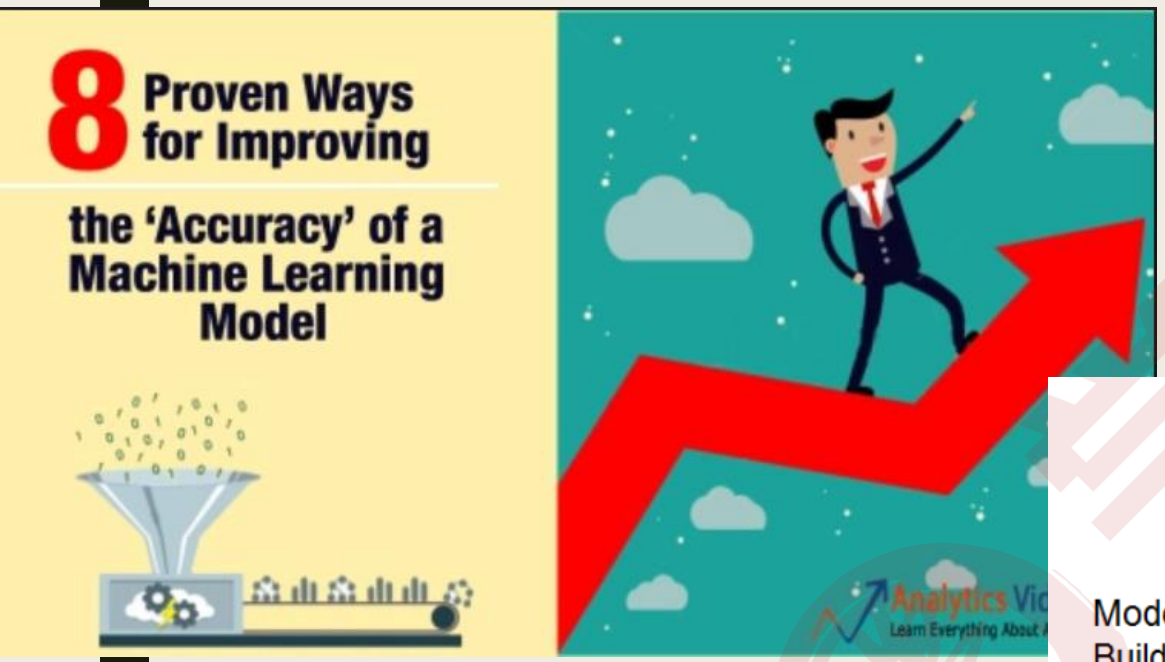
机器学习：测试模型的预测

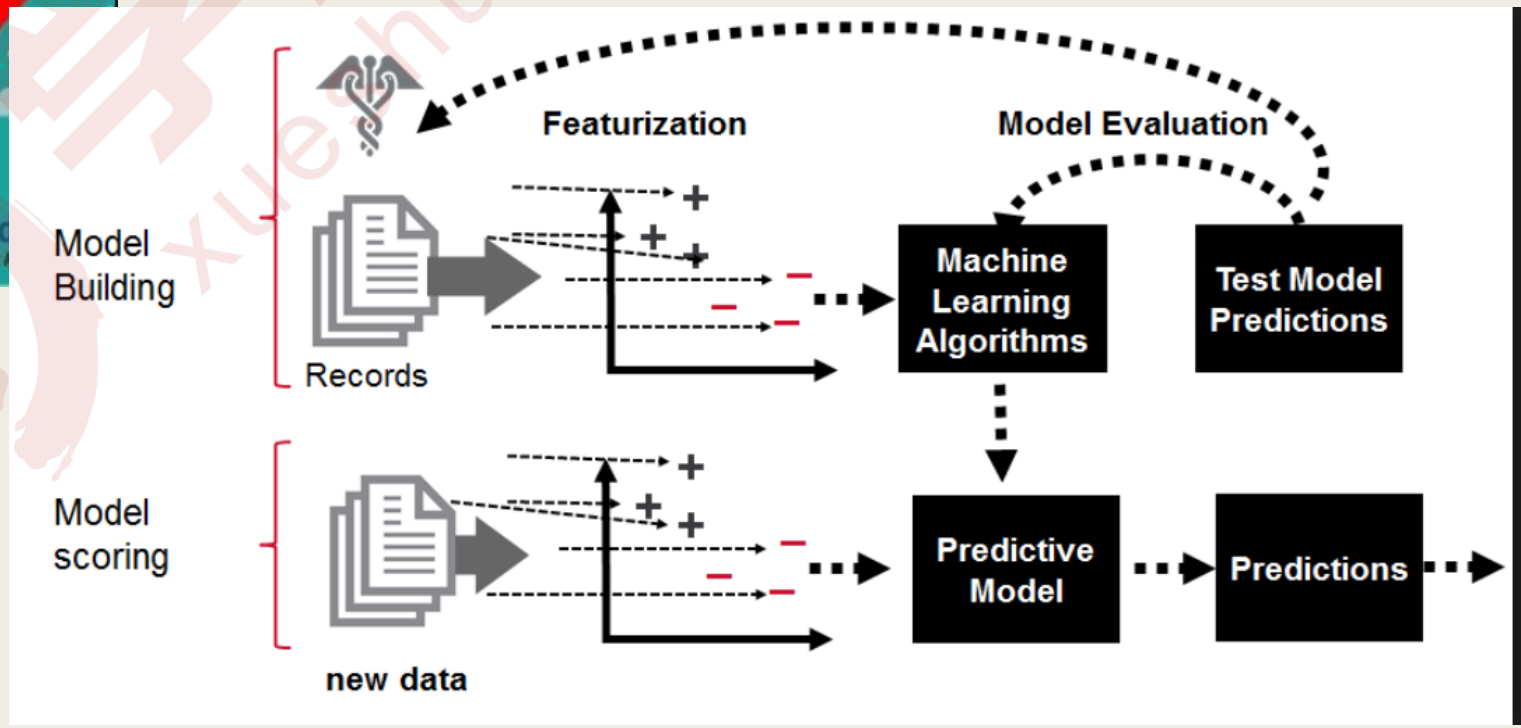


机器学习：模型评估


Test data
Training data
Iteration 1
Iteration 2
Iteration 3
Iteration k=4
All data

机器学习：模型改进


8 Proven Ways for Improving
the 'Accuracy' of a Machine Learning Model


Featurization
Model Evaluation
Model Building
Records
Machine Learning Algorithms
Test Model Predictions
Model scoring
new data
Predictive Model
Predictions

文本分析高阶

文本分类和聚类

- Grimmer(2013)文本分析的核心工作是分类（classification）。分类有三种方法：
  - 字典法（dictionary methods），根据关键词的出现次数来确定；
  - 有监督学习法（supervised learning methods)
  - 无监督学习法（unsupervised learning methods）

文本分类和聚类的应用

- 垃圾邮件识别（垃圾邮件/正常邮件：0/1）
- 新闻自动推送（政治、文化、娱乐等多分类）
- 情感分类
- 互联网舆情分析

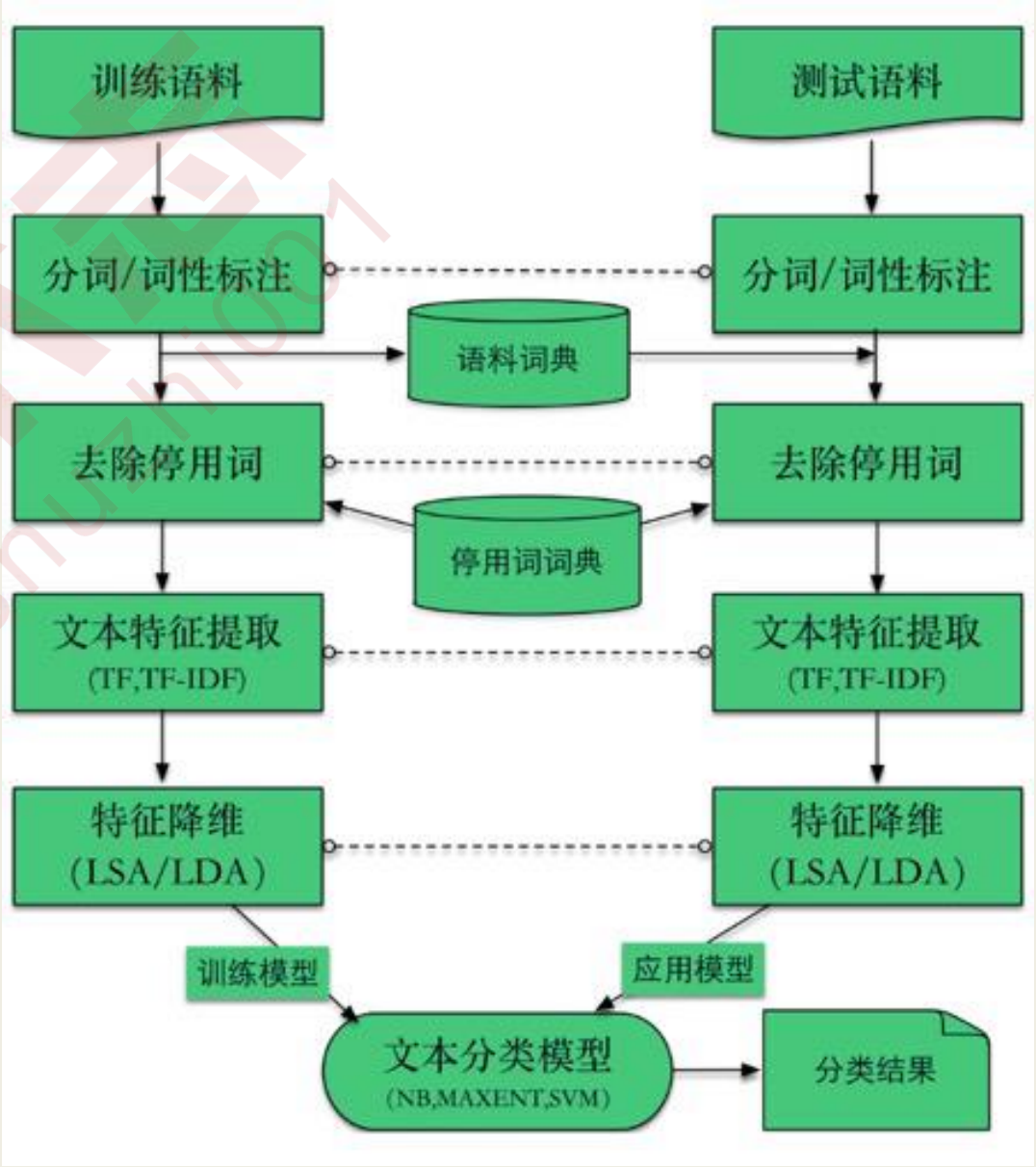
文本分类流程

与一般的分类的区别：

- 分词、词性标注、去除停用词等预处理流程
- 文本数据的结构化
- 文本降维

常见的分类算法

- Na ïve Bayes
- Support Vector Machines
- Maximum Entropy
- Random Forest

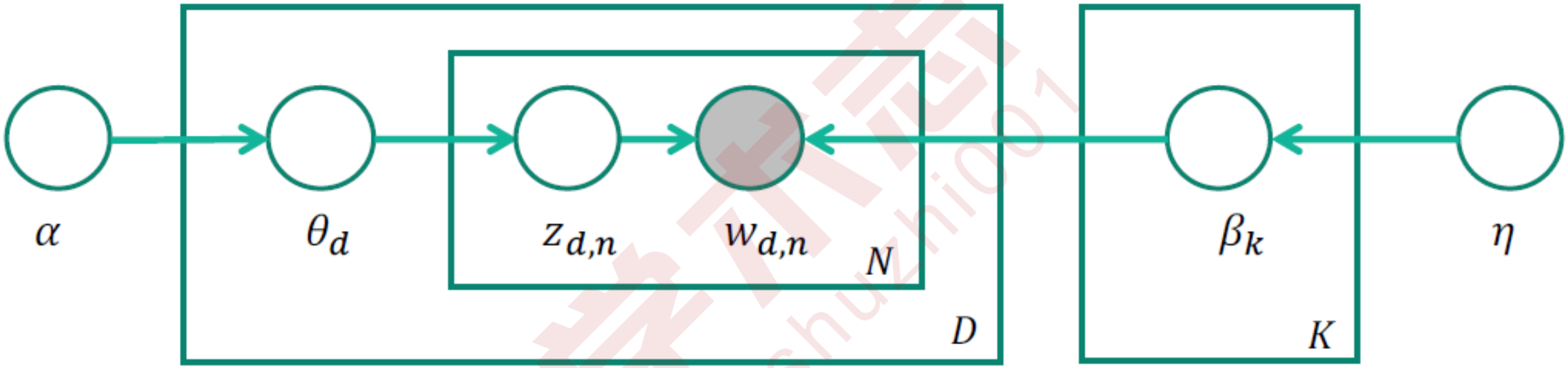


主题模型

- LDA模型由Blei, David M.、Ng, Andrew Y. 于2003年提出，是一种主题模型，它可以将文档集中每篇文档的主题以概率分布的形式给出，从而通过分析一些文档抽取出它们的主题（分布）出来后，便可以根据主题（分布）进行主题聚类或文本分类。同时，它是一种典型的词袋模型，即一篇文档是由一组词构成，词与词之间没有先后顺序的关系。
- 目的：根据给定的一篇文档，推测其主题分布。

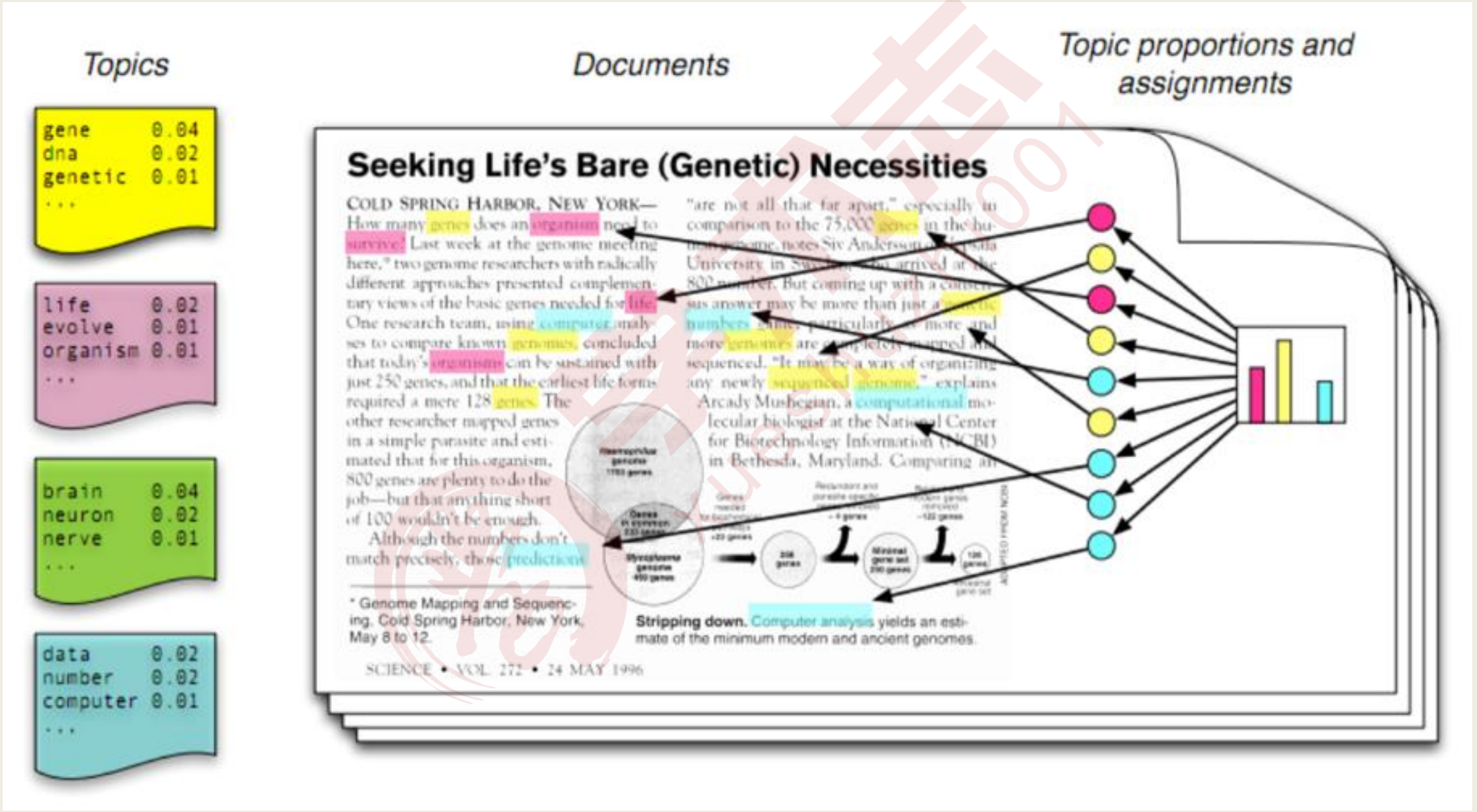
- 人：根据主题遣词造句，写成了各种各样的文章
- 计算机：推测分析网络上各篇文章分别都写了些啥主题，且各篇文章中各个主题出现的概率大小（主题分布）是啥

LDA目标



- 对于给定的文档集，推理：
  - 每一个词属于哪一个主题$z_{d,n}$
  - 每一篇文档的主题分布θ_d
  - 整个文档集中的主题分布β_k

LDA生成过程

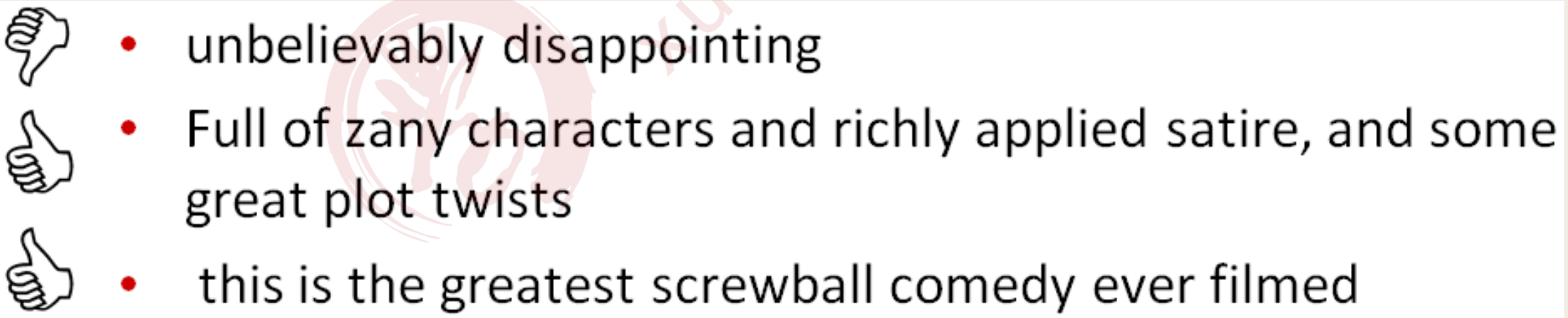

Topics
gene 0.04
dna 0.02
genetic 0.01
...
life 0.02
evolve 0.01
organism 0.01
...
brain 0.04
neuron 0.02
nerve 0.01
...
data 0.02
number 0.02
computer 0.01
...
Documents
Topic proportions and assignments
Seeking Life's Bare (Genetic) Necessities
COLD SPRING HARBOR, NEW YORK—
How many genes does an organism need to survive? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for life. One research team, using computer analyses to compare known genomes, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 genes. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.
Although the numbers don't match precisely, those predictions
"are not all that far apart," especially in comparison to the 75,000 genes in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a matter of numbers. Since particularly more and more genomes are completely mapped and sequenced, "it may be a way of organizing any newly sequenced genome," explains Arcady Mushegian, a computational molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an
* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.
SCIENCE • VOL. 272 • 24 MAY 1996

有监督机器学习

- 监督学习：通过已有的一部分输入数据与输出数据之间的相应关系。生成一个函数，将输入映射到合适的输出，比如分类。
- 有监督的机器学习应用到文本分析中，即将按照人工分类的标准输入给系统，预测尚未分类的文本的类别。其中，人工标注的文本作为特征，预测尚未分类的文本的类别归属。

情感分析

- 情感分析（Sentiment analysis），又称倾向性分析，意见抽取（Opinion extraction），意见挖掘（Opinion mining），情感挖掘（Sentiment mining），主观分析（Subjectivity analysis），它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。



情感分析的重要性

- 情感对人们行为的重要性
- 情感在人们行为中的重要性
- 情感分析的现实价值

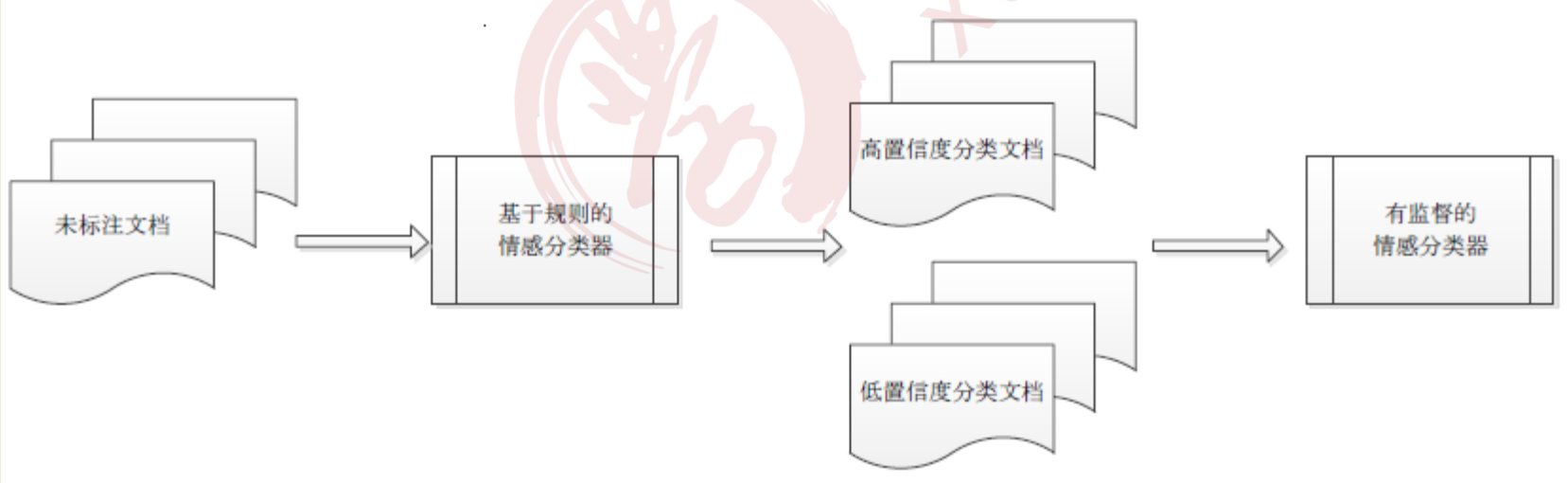
- 2012年5月，世界首家基于社交媒体的对冲基金 Derwent Capital Markets 在屡次跳票后终于上线。它会即时关注Twitter 中的公众情绪指导投资——1.85%的收益率，让平均数只有0.76%的其他对冲基金相形见绌
- “冷静CLAM”情绪指数后移3天后和道琼斯工业平均指数DIJA惊人一致（Johan Bollen, Huina Mao, Xiaojun Zeng. 2011. Twitter mood predicts the stock market, Journal of Computational Science 2:1, 1-8）

情感分析的分析内容

- 主客观分类（有监督机器学习）
  - 华为手机Mate10是今年推出的产品。
  - 我很喜欢华为手机Mate10
- 情感倾向性
  - 二类三类或者多类，甚至连续值
- 文档情感分类
  - 整体倾向，假设每篇文档仅针对一个主题的观念。
- 基于特性的情感分析
- 情感词典构建

情感分析的实现途径

- 情感词典
- 有监督方式
  - 将情感倾向性作为分类问题处理，主要在于特征选择
- 无监督方式
- 半监督方式



案例：人民网地方政府留言板

案例：新闻数据库

- 20 newsgroups数据集18000篇新闻文章，一共涉及到20种话题，所以称作20 newsgroups text dataset，分文两部分：训练集和测试集，通常用来做文本分类。一些新闻组的主题特别相似(e.g. comp.sys.ibm.pc.hardware/ comp.sys.mac.hardware)，还有一些却完全不相关 (e.g misc.forsale /soc.religion.christian)。
- sklearn提供了该数据的接口：sklearn.datasets.fetch_20newsgroups，我们以sklearn的文档来解释下如何使用该数据集。
- from sklearn.datasets import fetch_20newsgroups

谢谢！