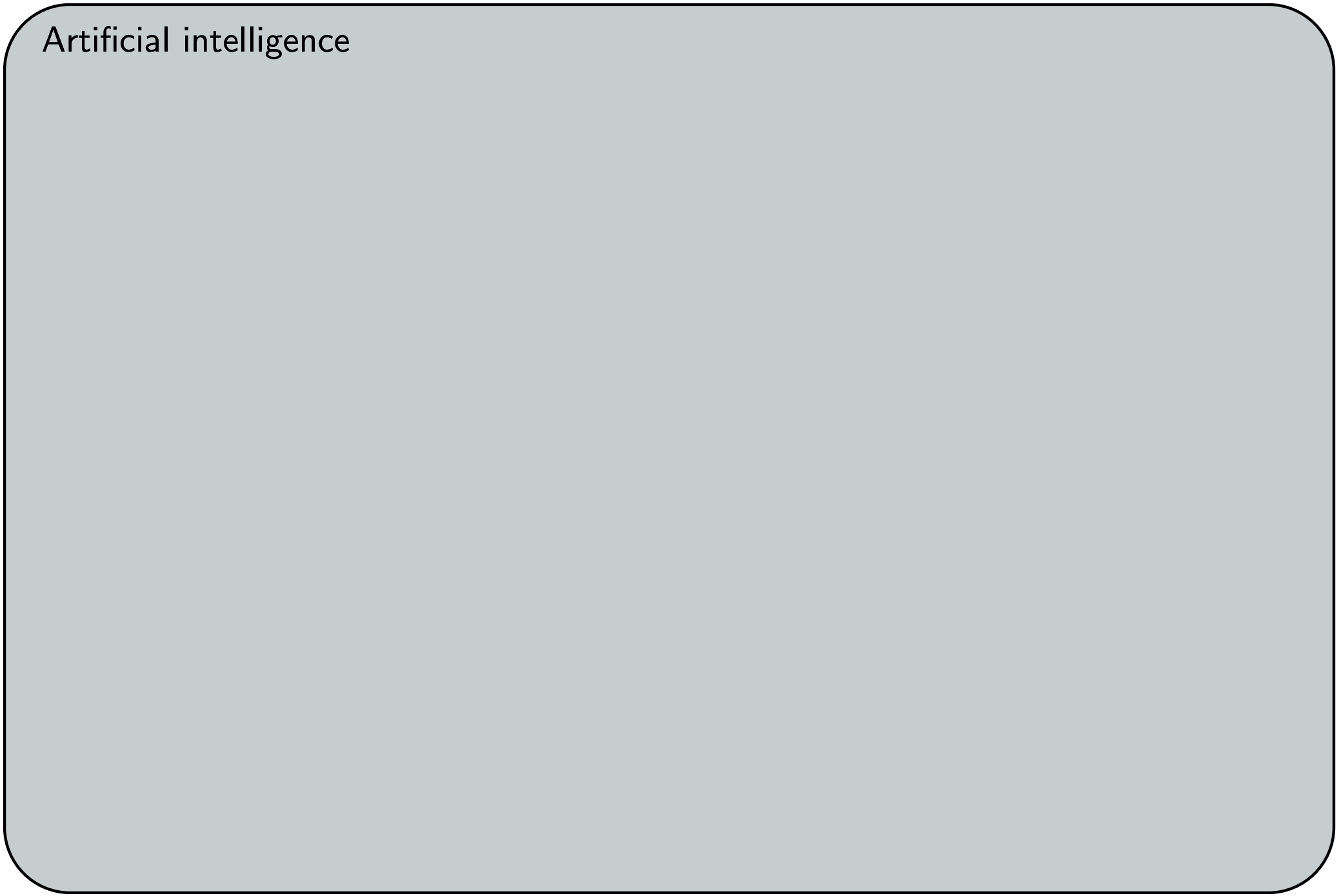
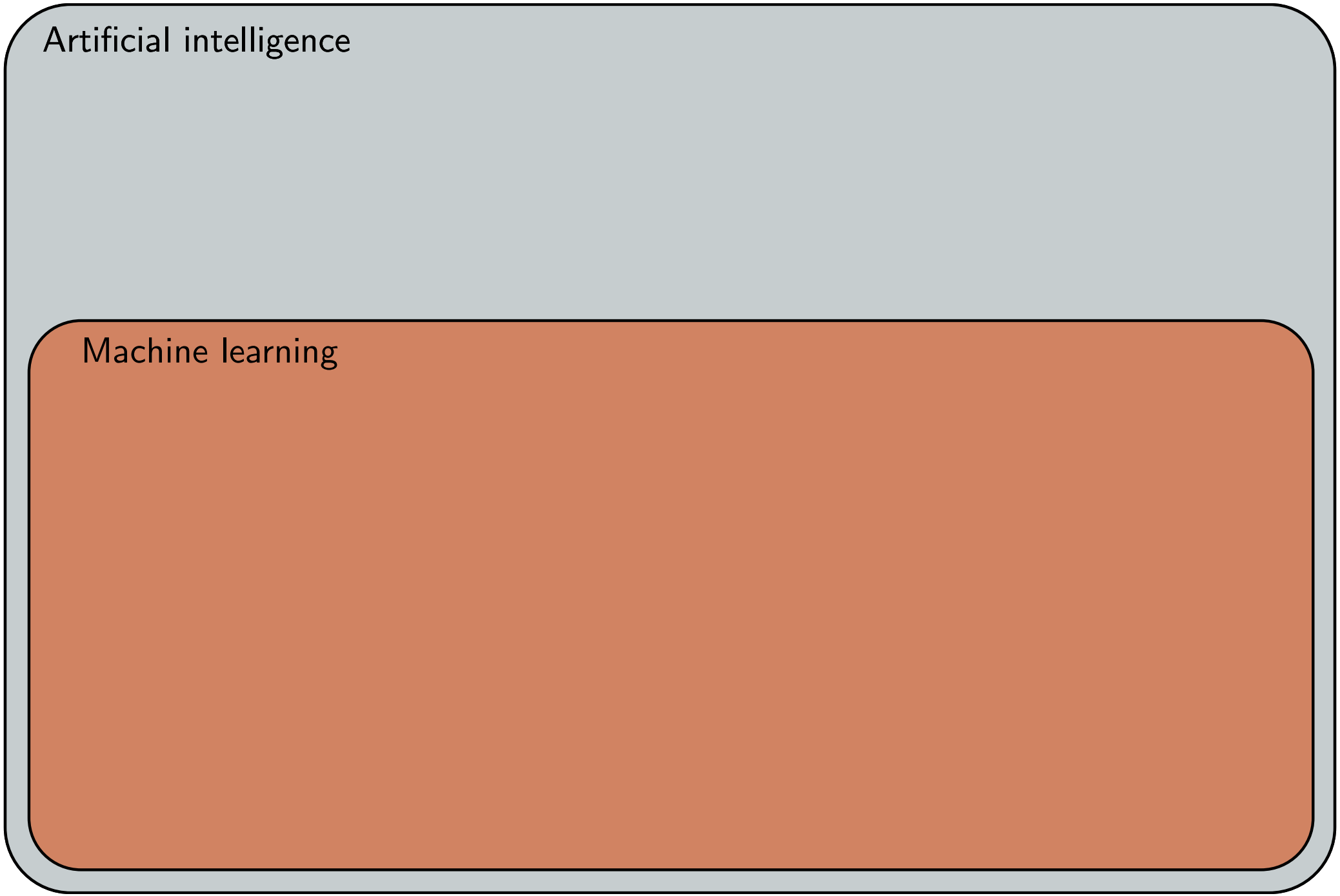
数据挖掘与机器学习

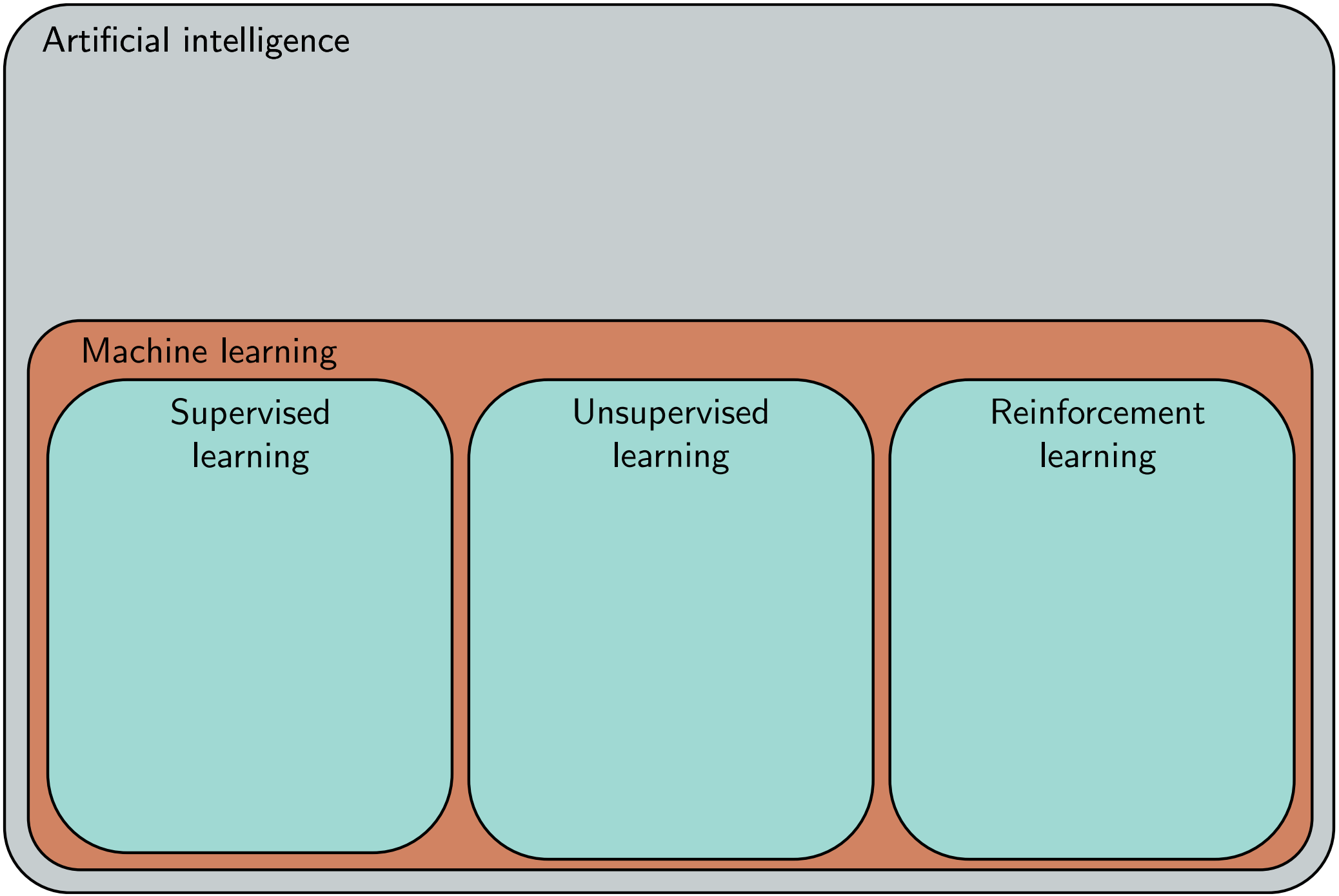
什么是机器学习（Machine learning）

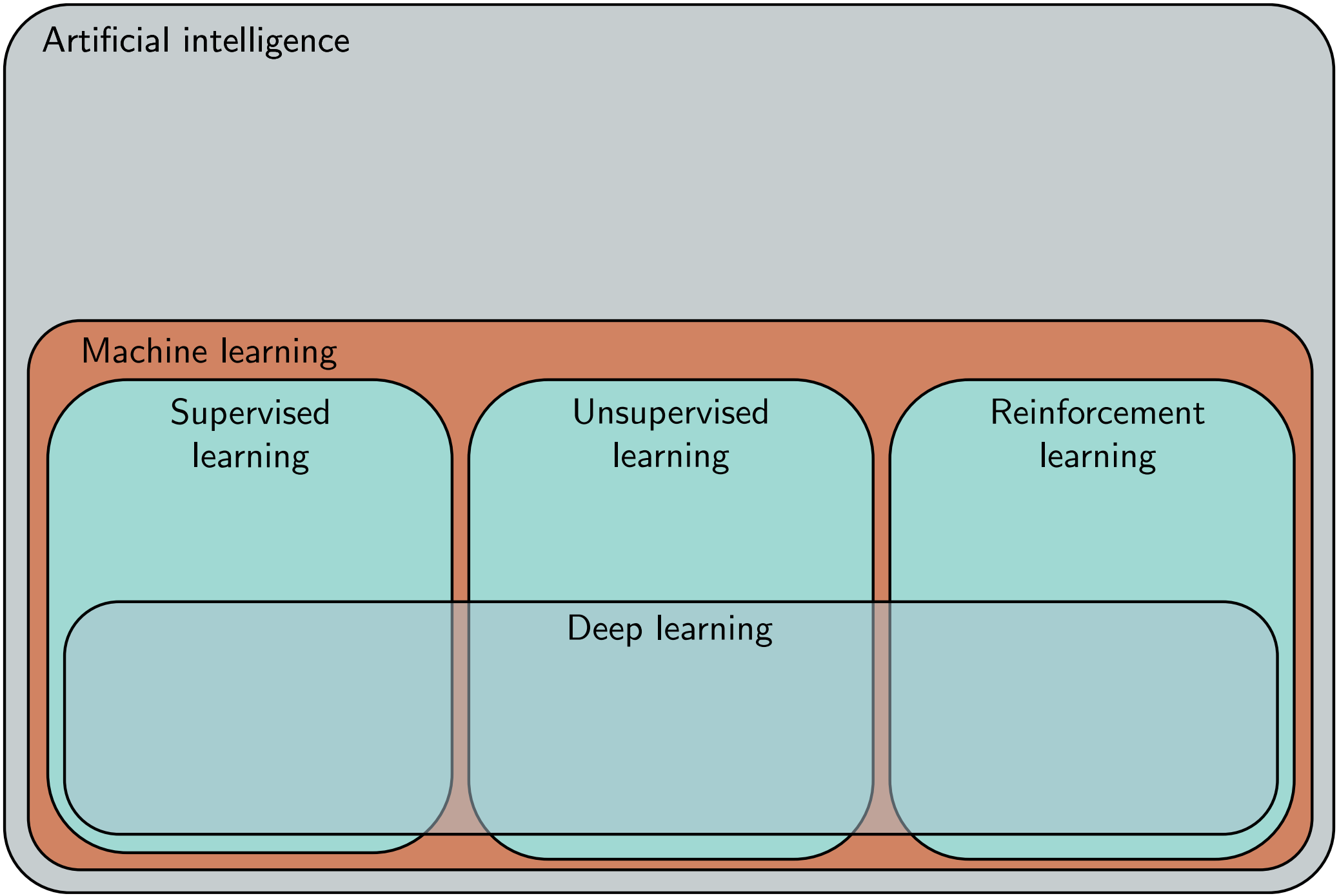
什么是机器学习？

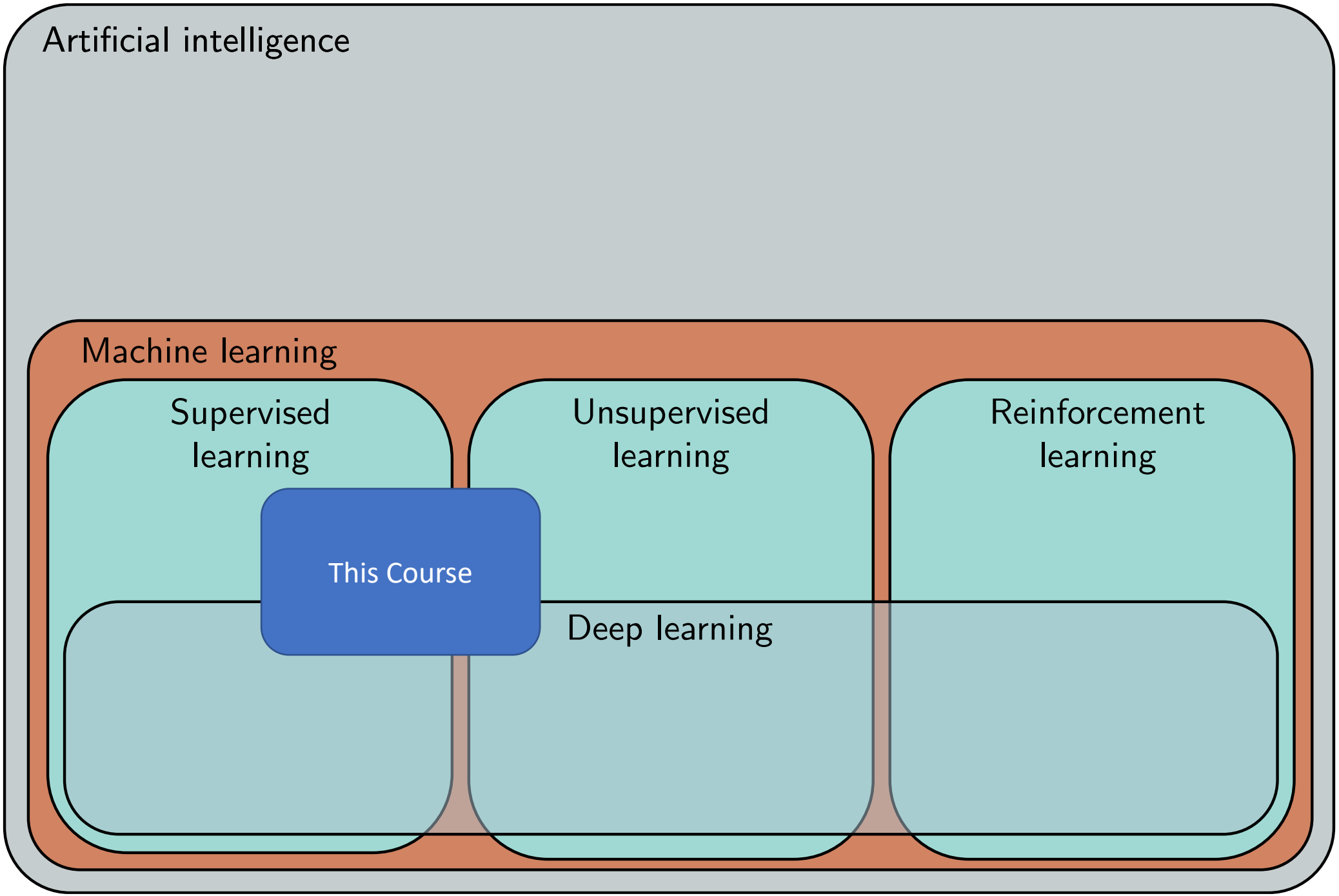
- **定义与核心思想**：从数据中学习模式，进行预测和决策。
  - 机器学习使计算机无需显式编程即可学习。
  - 核心在于从经验（数据）中自动改进。
- **与传统编程的区别**：
  - 传统编程：明确规则，处理确定性问题。
  - 机器学习：从数据中发现规则，处理不确定性和复杂性问题。
- **主要类型**：
  - **监督学习** (Supervised Learning)：有标签数据，预测结果。
  - **无监督学习** (Unsupervised Learning)：无标签数据，发现模式。
  - **强化学习** (Reinforcement Learning)：通过试错学习策略，优化。


Artificial intelligence


Artificial intelligence
Machine learning


Artificial intelligence
Machine learning
Supervised learning
Unsupervised learning
Reinforcement learning


Artificial intelligence
Machine learning
Supervised learning
Unsupervised learning
Reinforcement learning
Deep learning


Artificial intelligence
Machine learning
Supervised learning
Unsupervised learning
Reinforcement learning
This Course
Deep learning

课程目标与考核方式

- 课程目标
  - **理论**：掌握机器学习的基本理论和常用算法。
  - **实践**：熟悉Python数据分析和机器学习工具库 (Pandas, Scikit-learn等)。
  - **应用**：能够运用机器学习方法解决金融领域的实际问题。
  - **效率**：培养利用AI辅助工具进行高效编程和问题解决的能力 (Cursor, Copilot)。
- 考核方式
  - **综合项目1 (25%)**：借贷违约风险评估模型
  - 特征工程 + 模型评估报告 + 复现代码
  - **综合项目2 (25%)**：股票价格预测/财报文本分析
  - 机器学习部分 + 金融预测部分 + 复现代码
  - **期末考试 (50%)**：闭卷笔试
  - 理论80% + 案例分析20%

综合项目1：借贷违约风险评估模型

- **项目背景**：
- P2P借贷平台风险管理至关重要，降低坏账率是平台生存和发展的关键。
- **项目目标**：
- 利用Lending Club数据集，构建机器学习模型预测借款人是否违约。
- **Lending Club数据集**：
- 来源：美国P2P借贷平台 Lending Club 公开数据。
- 规模：包含大量借款人的个人和贷款信息。
- 特征字段：借款人特征 (年龄、收入、信用评分等)，贷款特征 (贷款金额、利率、期限等)。
- **评价指标**：
- 准确率 (Accuracy), 精确率 (Precision), 召回率 (Recall), F1-score, AUC等。
- 强调根据业务场景选择合适的评估指标 (例如，关注坏账率，则Recall更重要)。

综合项目2A：股票价格预测

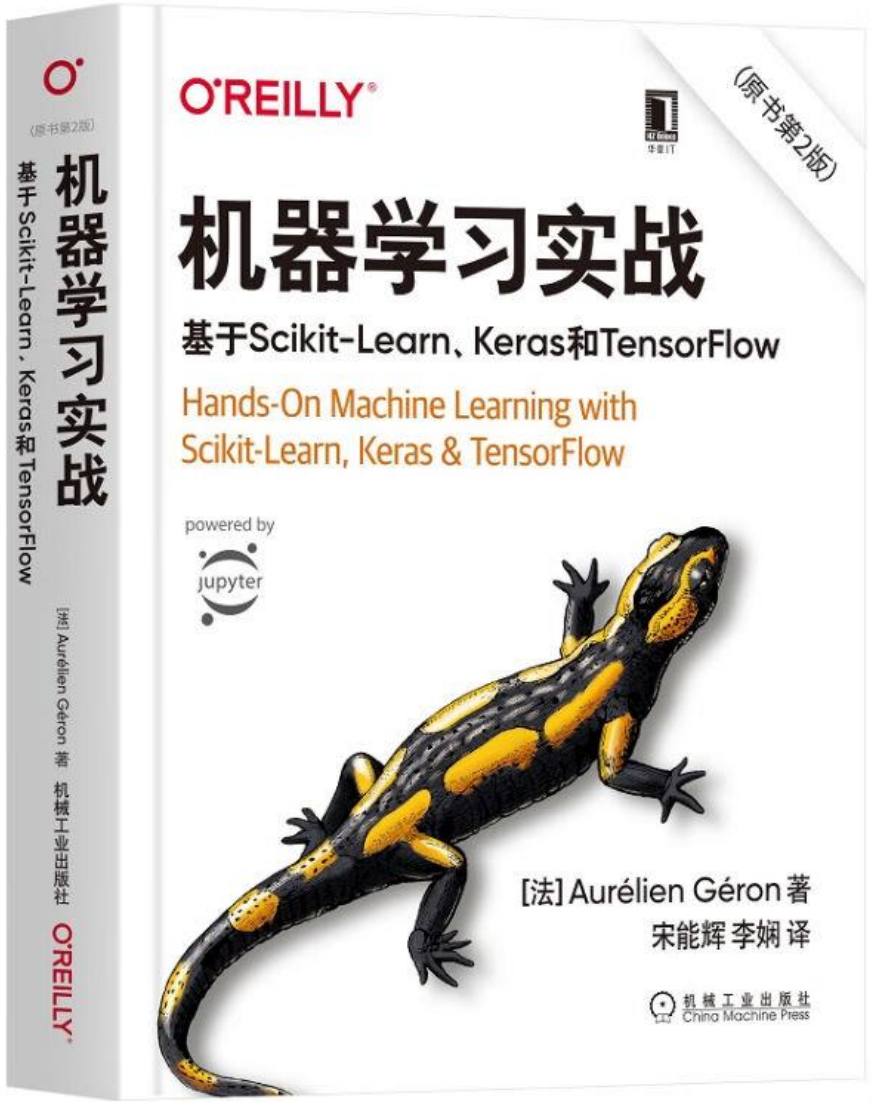
- **项目背景**：
- 量化投资依赖于对未来市场走势的预测。
- 机器学习为股票价格预测提供了新的工具和方法。
- **项目目标**：
- 利用历史股票数据，构建机器学习模型预测未来股票价格走势。
- 掌握时间序列数据处理和机器学习模型应用。
- **数据获取与特点**：
- 数据来源：Tushare金融数据接口, CSMAR, WIND等。
- 时间序列数据特点：时间依赖性，自相关性，趋势性，季节性等。
- **评价指标**：
- 均方根误差 (RMSE), 平均绝对误差 (MAE), R-squared等。
- 根据预测目标选择合适的指标 (例如，关注预测精度，则RMSE更常用)。

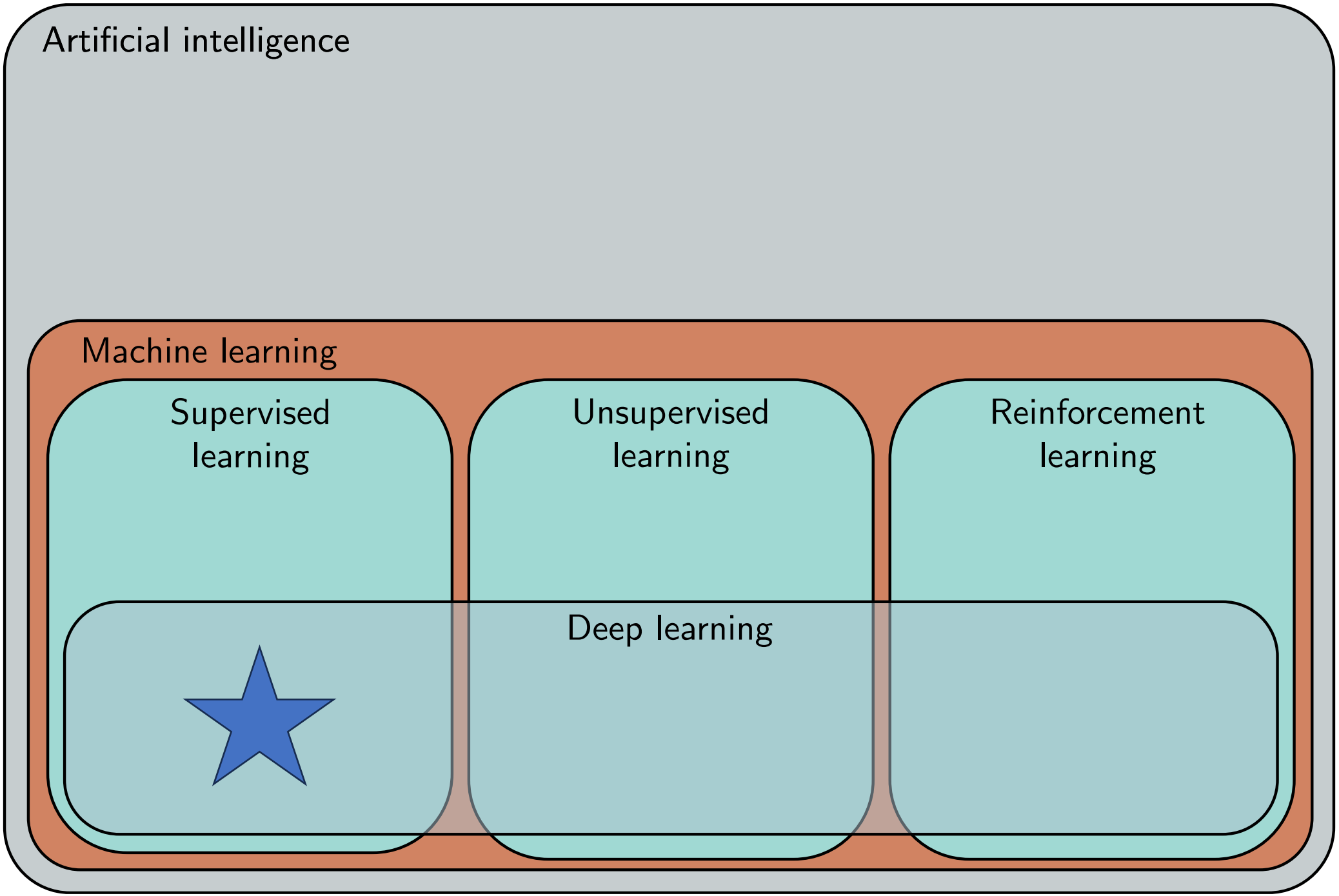
综合项目2B：财报文本分析

- **项目背景**:
- 财务报告包含大量非结构化文本信息。
- 文本信息蕴含重要的公司经营风险信息。
- **项目目标**:
- 利用上市公司财报数据，进行文本分析，提取关键信息。
- 运用机器学习方法辅助金融预测，例如：风险预警。
- **上市公司财报数据获取**:
- 数据来源：巨潮资讯网。
- **评价指标**:
- 准确率, 精确率, 召回率, F1-score等。
- 根据公司经营风险的定义选择合适的指标。

参考书目


机器学习
周志华 著
清华大学出版社
周志华著
MACHINE LEARNING
机器学习
清华大学出版社

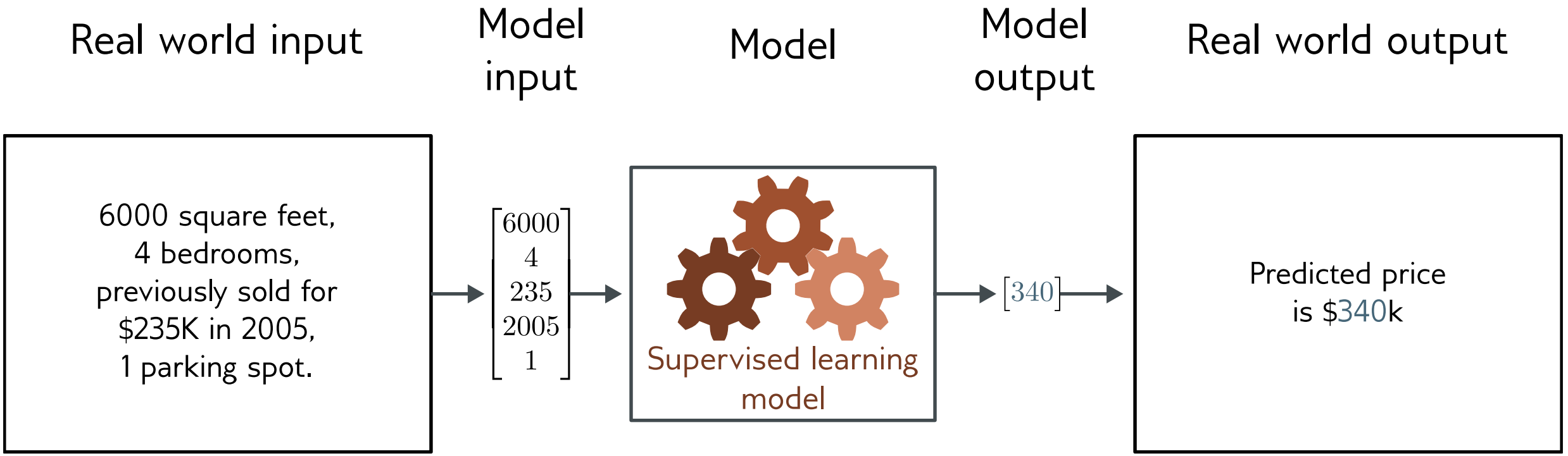

O'REILLY
(原书第2版)
机器学习实战
基于Scikit-Learn、Keras和TensorFlow
Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow
powered by
jupyter
[法] Aurélien Géron 著
宋能辉 李娴 译
机械工业出版社 China Machine Press
(原书第2版)
机器学习实战
基于Scikit-Learn、Keras和TensorFlow
[法] Aurélien Géron 著 机械工业出版社
O'REILLY


Artificial intelligence
Machine learning
Supervised learning
Unsupervised learning
Reinforcement learning
Deep learning

监督学习（Supervised learning）

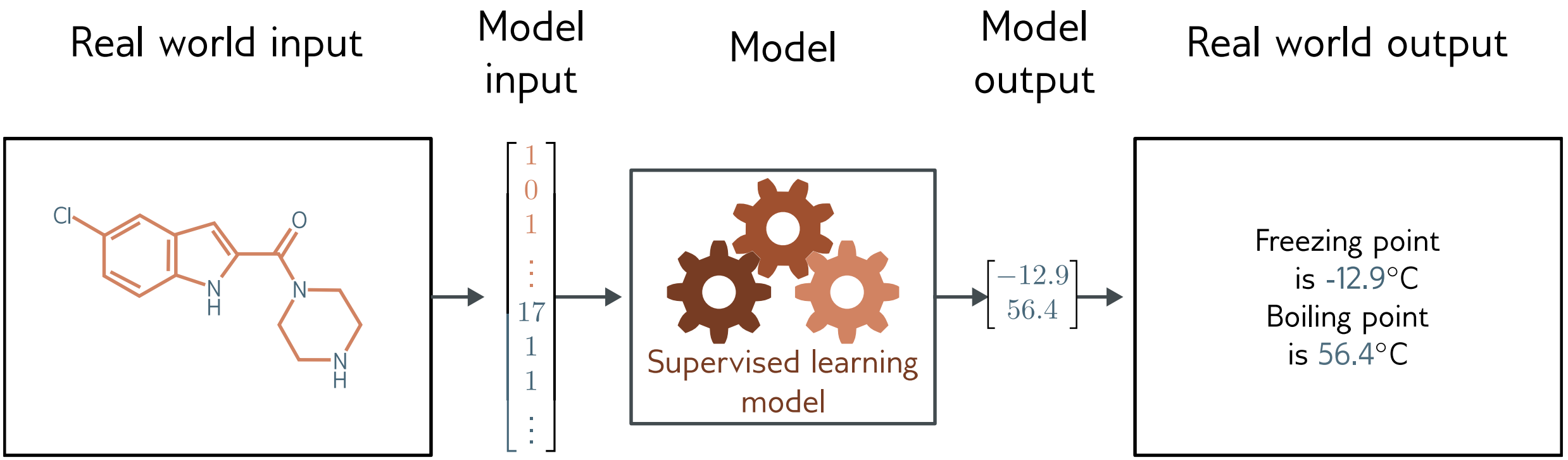
- 定义一个从输入到输出的映射
- 从输入-输出数据样例中学习这个映射的函数表达

回归（Regression）



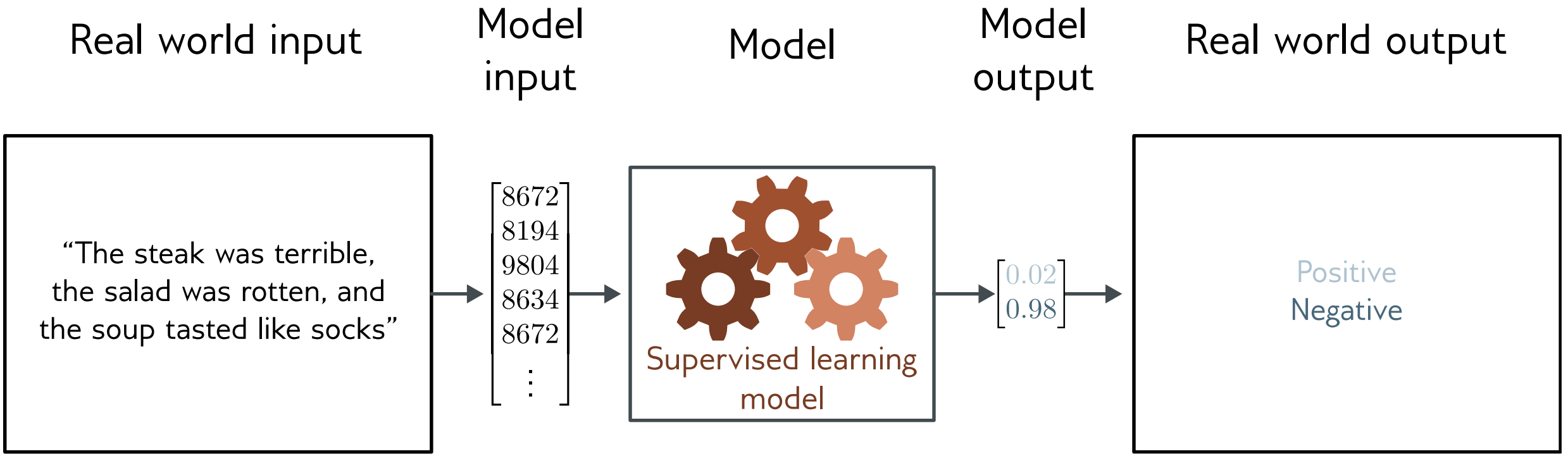
- 一元回归问题（单维、数值型输出）

图回归（Graph regression）



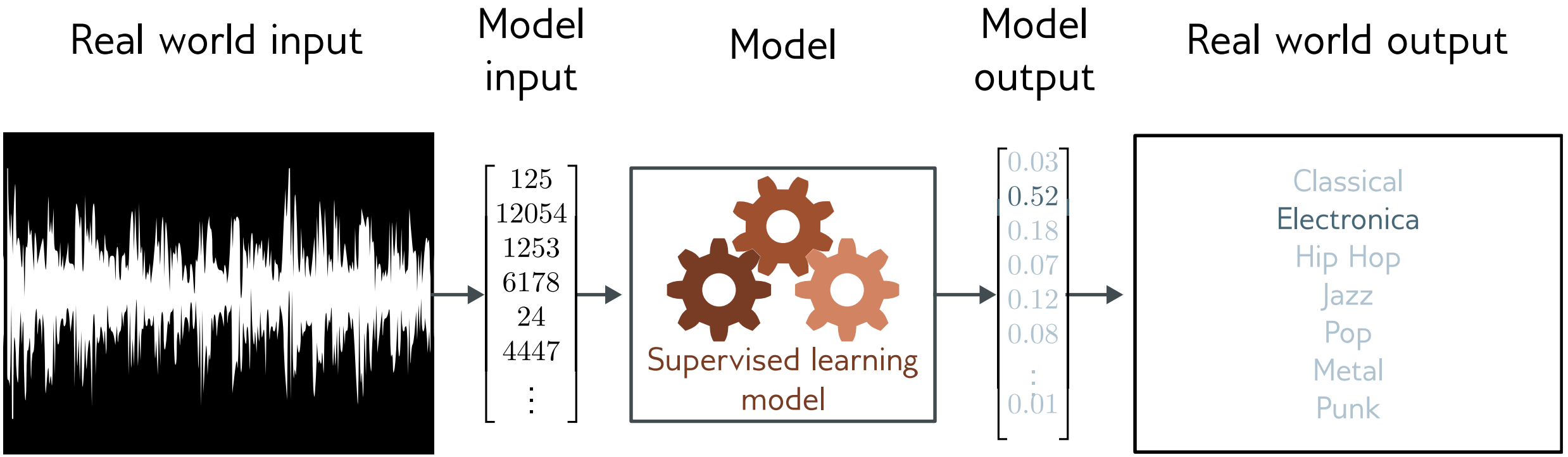
- 多元回归问题（多维、数值型输出）

文本分类（Text classification）



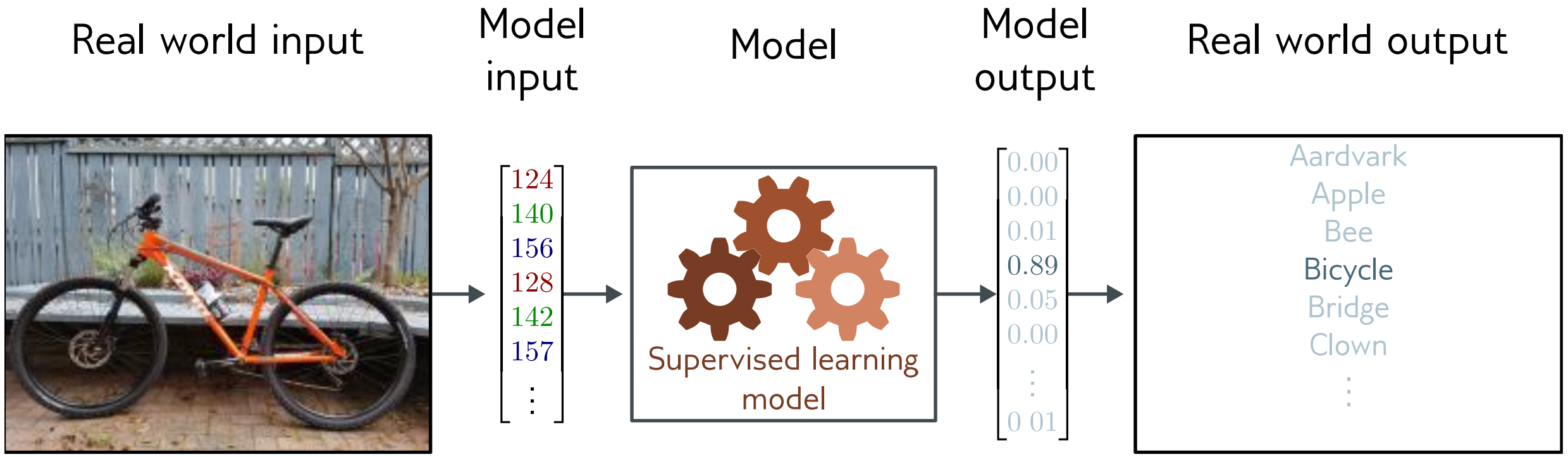
- 二分类问题（两个类别）

音乐类型分类（Music genre classification）



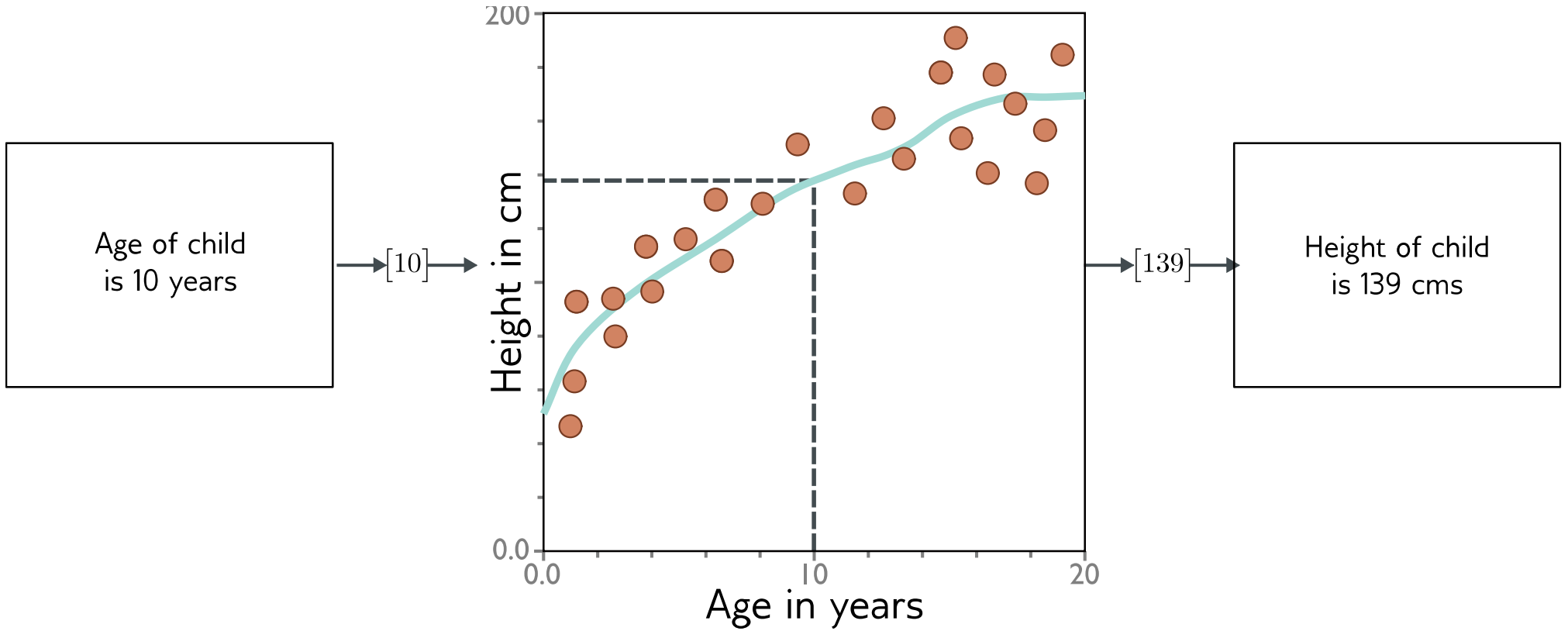
- 多分类问题（离散类别，大于等于两种取值）

图片分类（Image classification）



- 多分类问题（离散类别，大于等于两种取值）

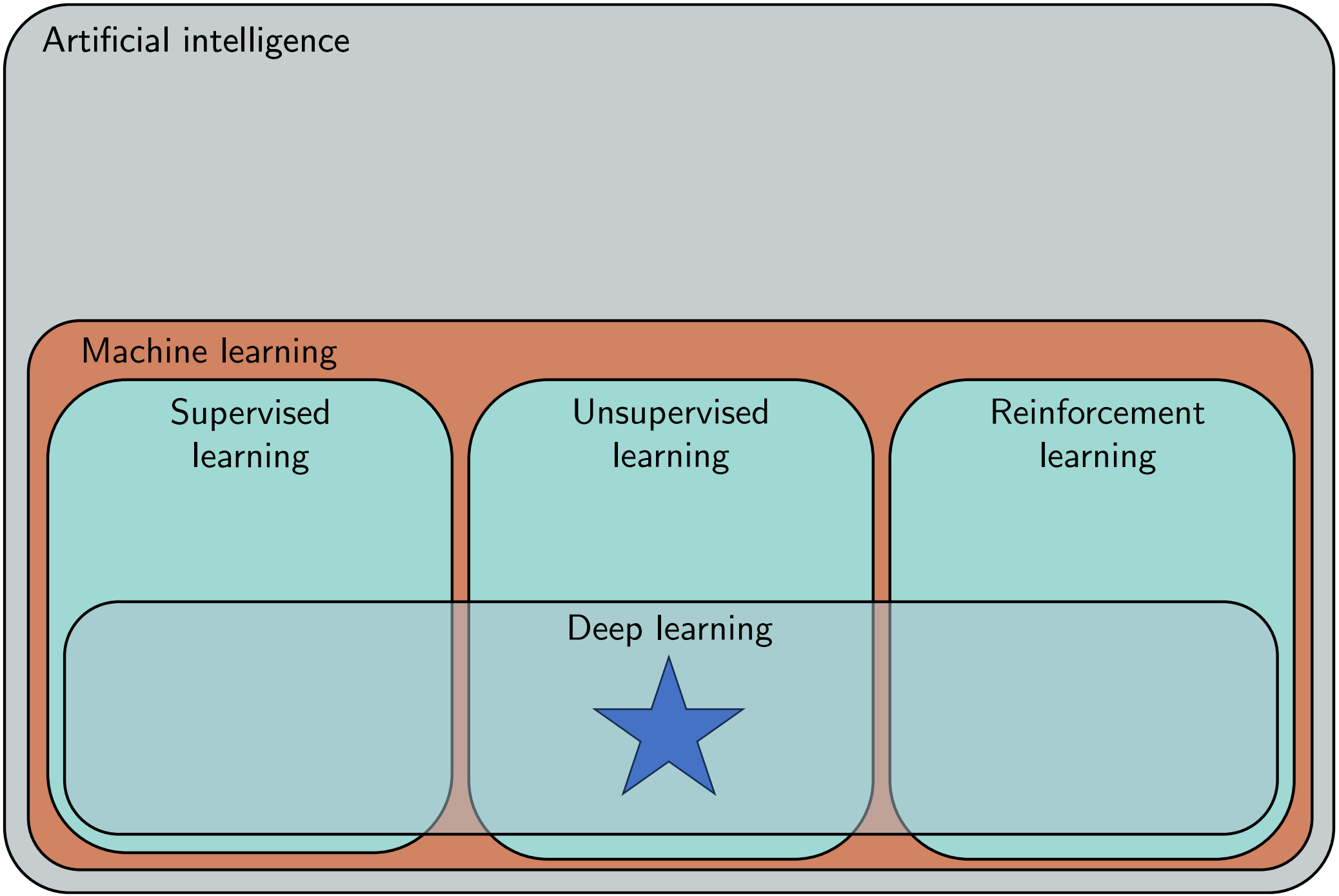
什么是一个监督学习模型？



- 一个把输入（年龄）与输出（身高）联系起来的方程
- 搜索可能的方程，找到可以把训练数据拟合得好的那个方程

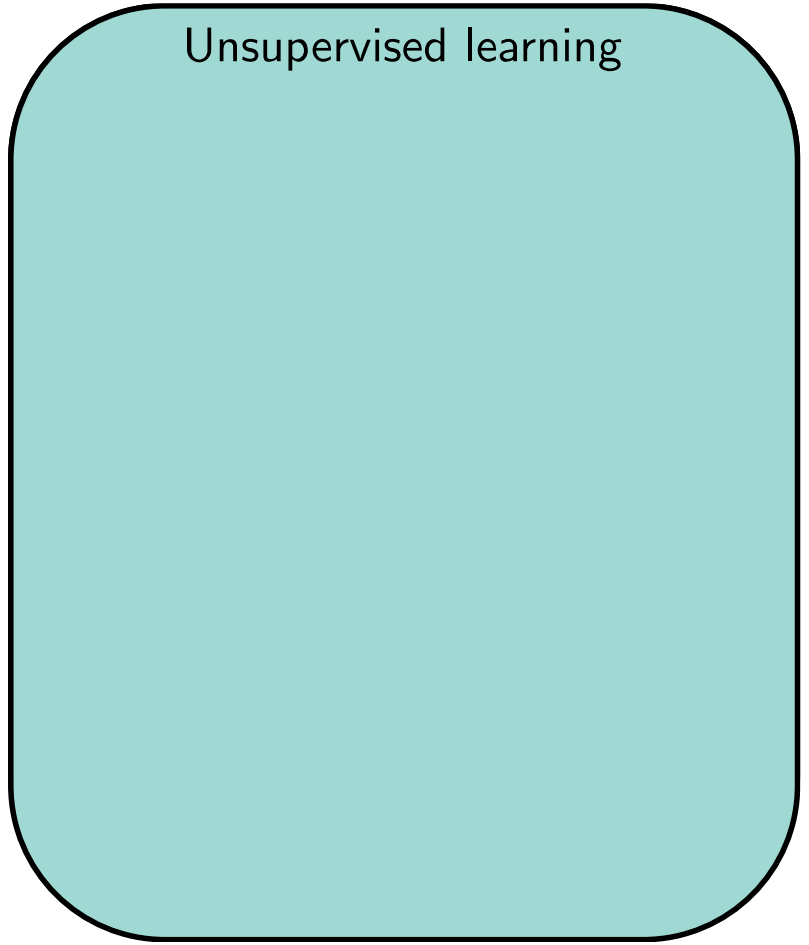
术语

- 回归（Regression） = 输出是连续型变量
- 分类（Classification） = 输出是分类型变量
- 二分类和多分类问题不同
- 一元（Univariate） = 一维输出
- 多元（Multivariate） = 多维输出

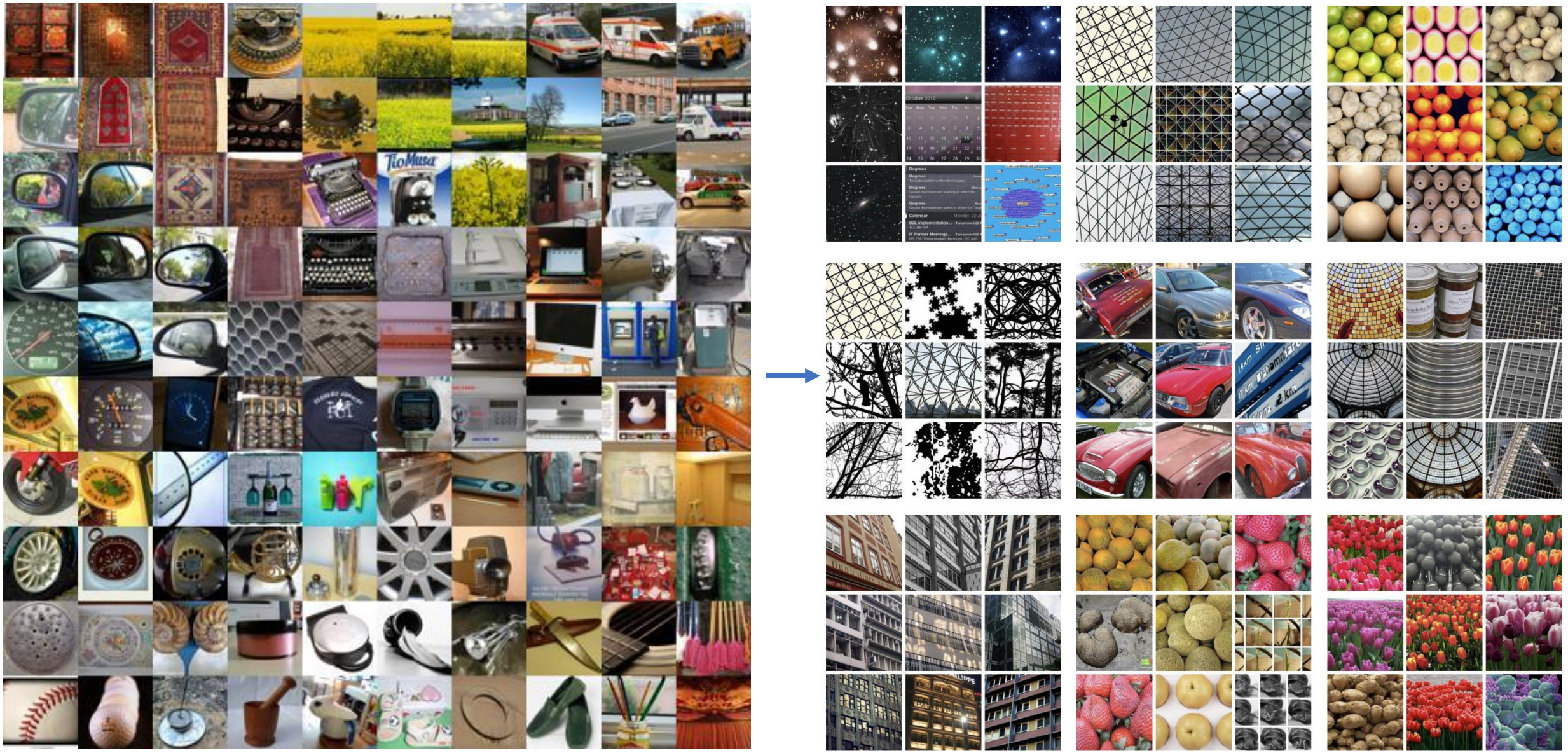

Artificial intelligence
Machine learning
Supervised learning
Unsupervised learning
Reinforcement learning
Deep learning

无监督学习（Unsupervised Learning）

- 对一个无标签的数据集进行学习
  - 聚类
  - 找出异常值
  - 生成新样例
  - 填补缺失值



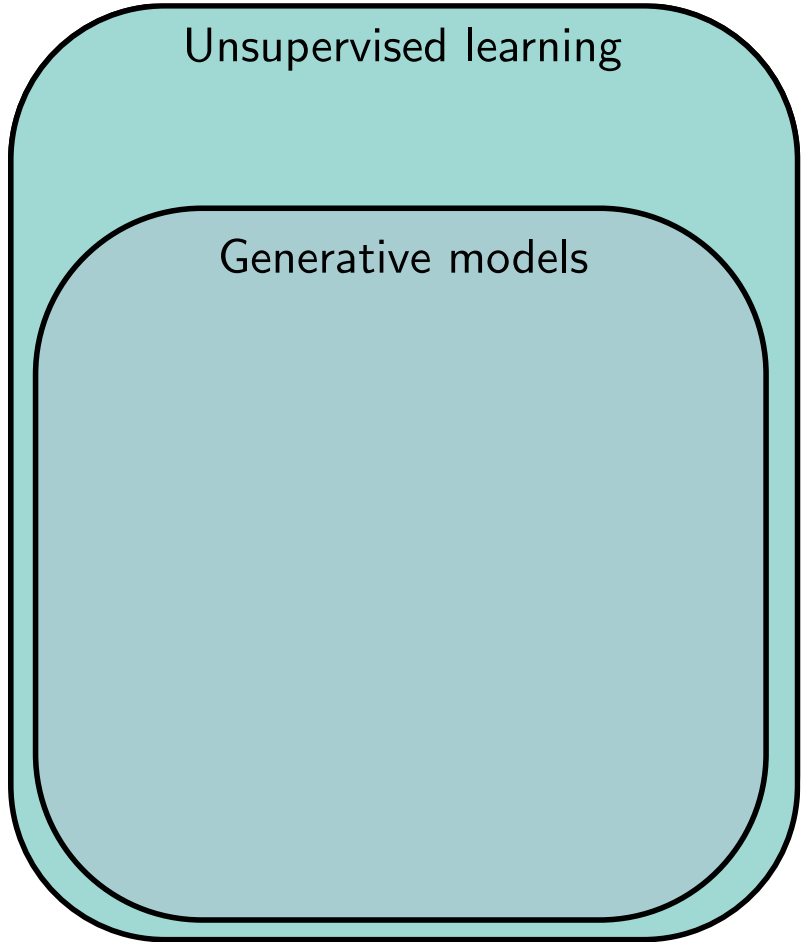
DeepCluster: Deep Clustering for Unsupervised Learning of Visual Features (Caron et al., 2018)



DeepCluster: Deep Clustering for Unsupervised Learning of Visual Features (Caron et al., 2018)

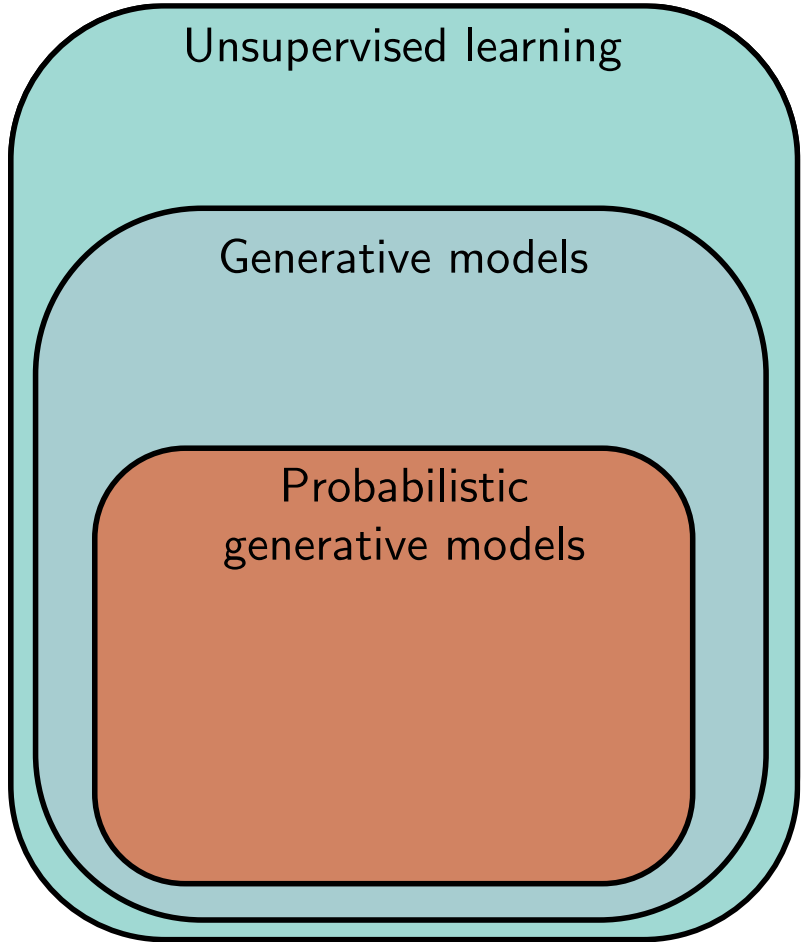
无监督学习

- 对一个无标签的数据集进行学习
  - e.g., 聚类
- 生成式模型可以生成新样例
  - e.g., 生成对抗网络（GANs）

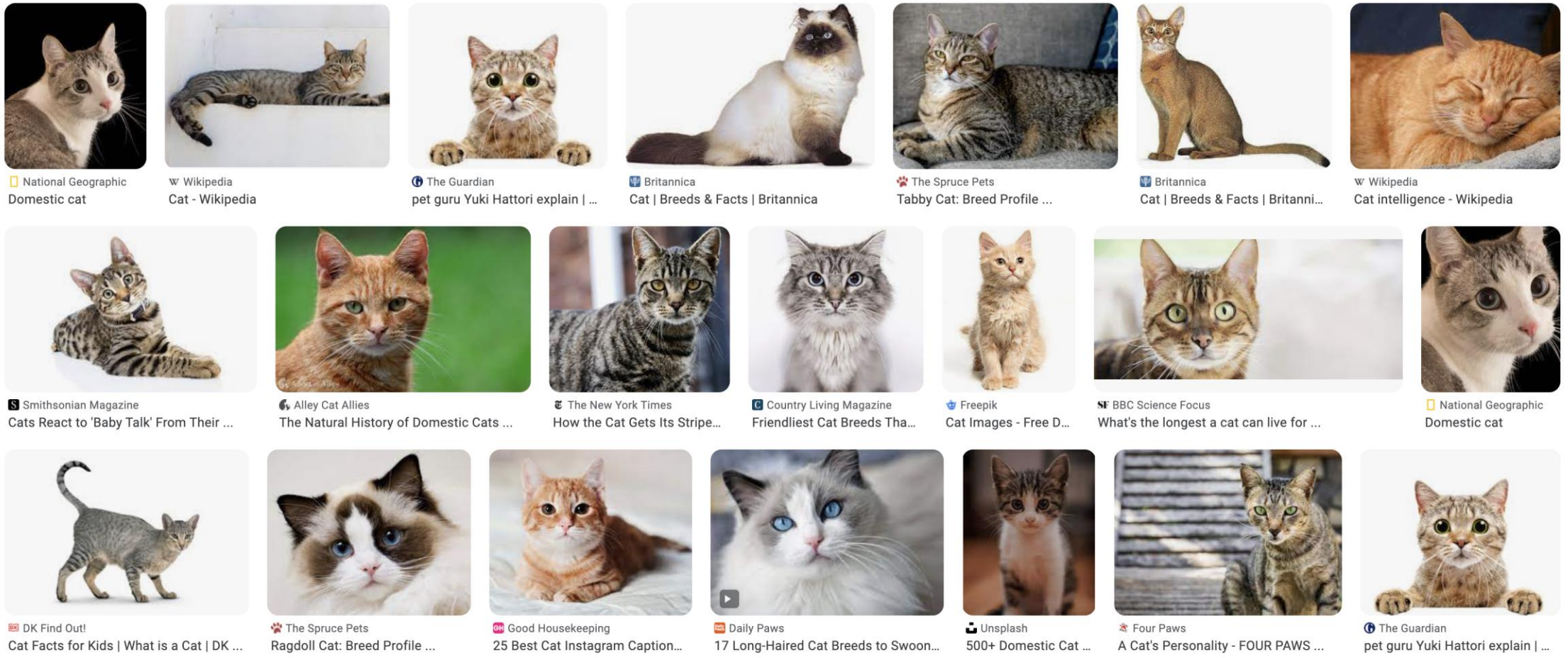


无监督学习

- 对一个无标签的数据集进行学习
  - e.g., 聚类
- 生成式模型可以生成新样例
  - e.g., 生成对抗网络（GANs）
- 概率生成模型通过数据学习分布
  - e.g., 变分自编码器（编码器+解码器），
  - e.g., 归一化流，
  - e.g., 扩散模型

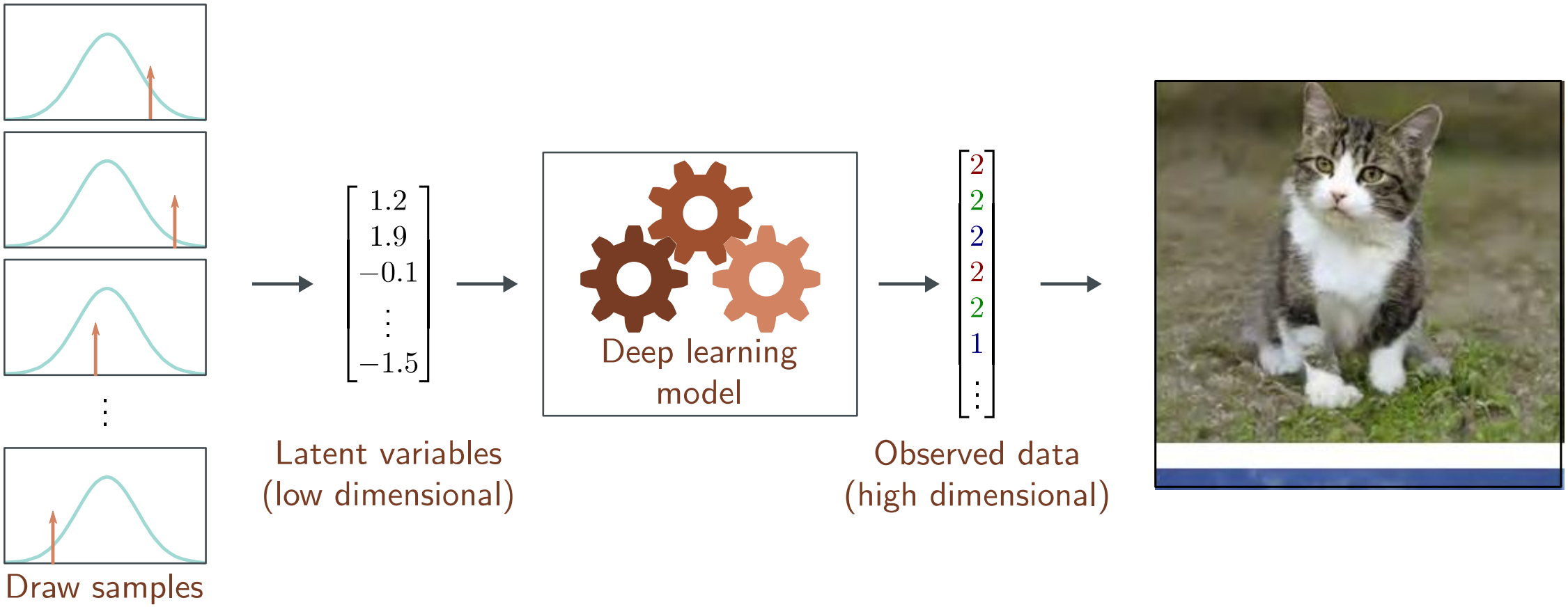


生成式模型-训练数据集


National Geographic
Domestic cat
Wikipedia
Cat - Wikipedia
The Guardian
pet guru Yuki Hattori explain | ...
Britannica
Cat | Breeds & Facts | Britannica
The Spruce Pets
Tabby Cat: Breed Profile ...
Britannica
Cat | Breeds & Facts | Britanni...
Wikipedia
Cat intelligence - Wikipedia
Smithsonian Magazine
Cats React to 'Baby Talk' From Their ...
Alley Cat Allies
The Natural History of Domestic Cats ...
The New York Times
How the Cat Gets Its Stripe...
Country Living Magazine
Friendliest Cat Breeds Tha...
Freepik
Cat Images - Free D...
BBC Science Focus
What's the longest a cat can live for ...
National Geographic
Domestic cat
DK Find Out!
Cat Facts for Kids | What is a Cat | DK ...
The Spruce Pets
Ragdoll Cat: Breed Profile ...
Good Housekeeping
25 Best Cat Instagram Caption...
Daily Paws
17 Long-Haired Cat Breeds to Swoon...
Unsplash
500+ Domestic Cat ...
Four Paws
A Cat's Personality - FOUR PAWS ...
The Guardian
pet guru Yuki Hattori explain | ...

生成式模型–输出

隐变量（Latent variables）



图像插值（Interpolation）

I was a little nervous before my first lecture at the University of Bath. It seemed like there were hundreds of students and they looked intimidating. I stepped up to the lectern and was about to speak, when something bizarre happened.

Suddenly, the room was filled with a deafening noise, like a giant roar. It was so loud that I couldn't hear anything else and I had to cover my ears. I could see the students looking around, confused and frightened. Then, as quickly as it had started, the noise stopped and the room was silent again.

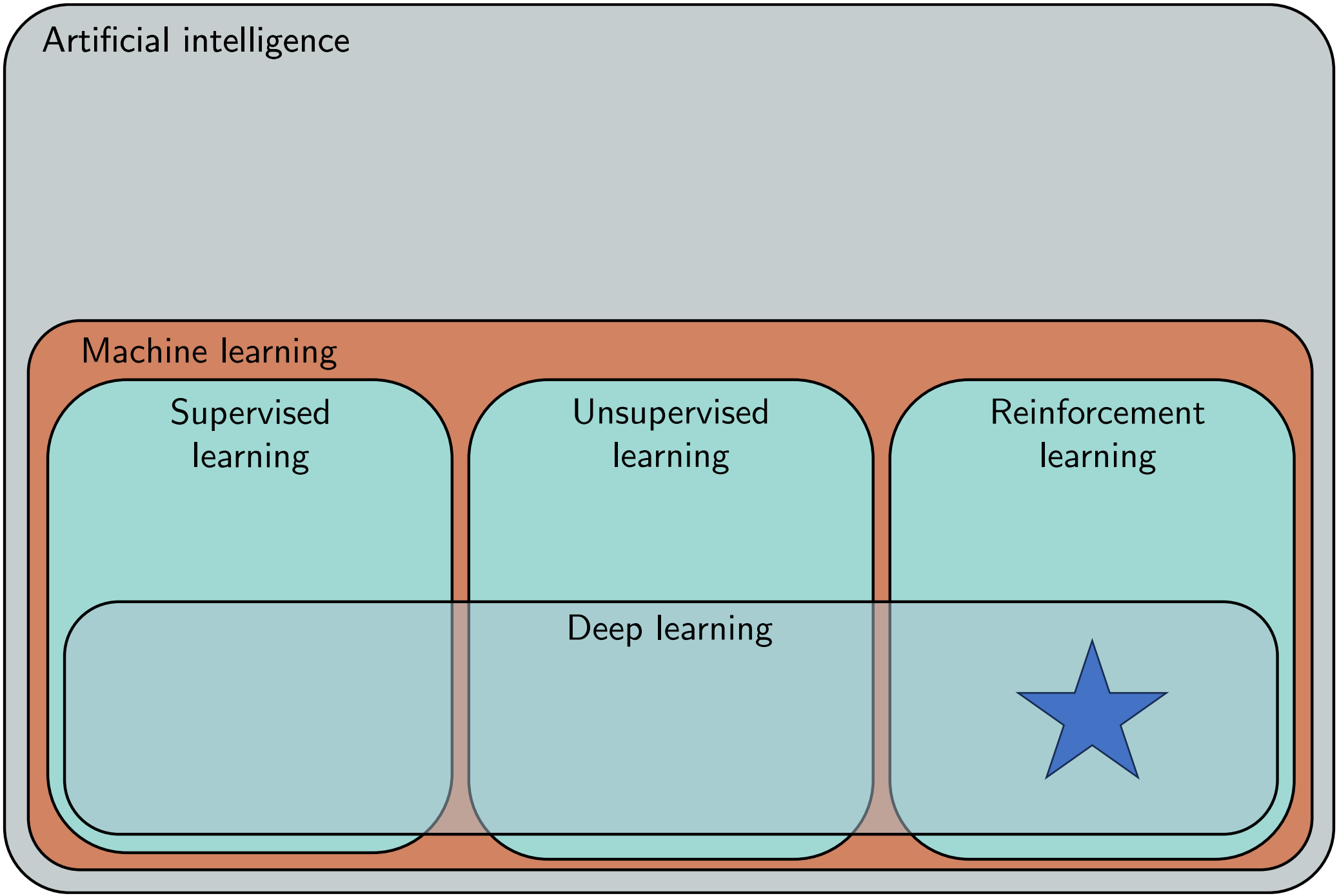
I stood there for a few moments, trying to make sense of what had just happened. Then I realized that the students were all staring at me, waiting for me to say something. I tried to think of something witty or clever to say, but my mind was blank. So I just said, "Well, that was strange," and then I started my lecture.

I was a little nervous before my first lecture at the University of Bath. It seemed like there were hundreds of students and they looked intimidating. I stepped up to the lectern and was about to speak, when something bizarre happened.

Suddenly, a giant rabbit ran into the lecture hall! The students started screaming and running around in panic. I was so shocked that I couldn't move. The rabbit ran up to me and hopped onto the lectern. Then, in a booming voice, it said:

"I am the Easter Bunny! I have come to give you all a special gift!"

The students were so surprised that they stopped screaming and listened to the Easter Bunny. Then, the Easter Bunny started handing out chocolate eggs to everyone in the lecture hall. The students were so happy that they started cheering and clapping. I was so relieved that the Easter Bunny had saved my lecture! After that, I was able to continue and the students paid attention for the rest of the hour. It was a great success!


Artificial intelligence
Machine learning
Supervised learning
Unsupervised learning
Reinforcement learning
Deep learning

强化学习（Reinforcement learning）

- 一组状态（states）
- 一组动作（actions）
- 一组奖励（rewards）

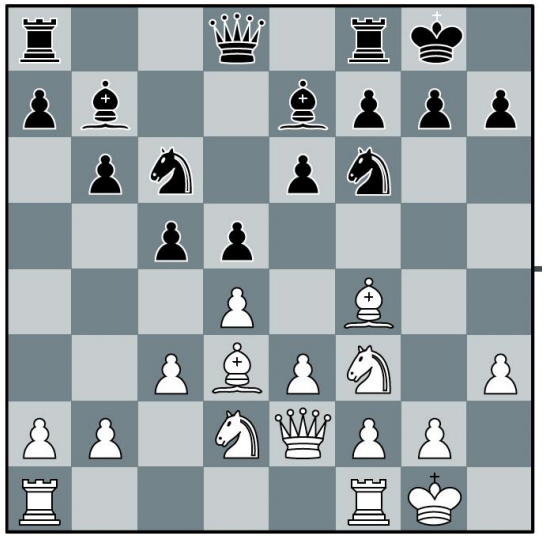
- 目标：采取行动（actions）改变状态（state）获得奖励（rewards）

- 没有现成的数据 – 通过自己摸索环境来采集数据

例子：国际象棋

- 状态是棋盘上棋子的当前布局
- 动作是当前符合规则的走棋方式
- 成功吃掉对方棋子可以得到正奖励，而自己的棋子被对方吃掉会得到负奖励

State



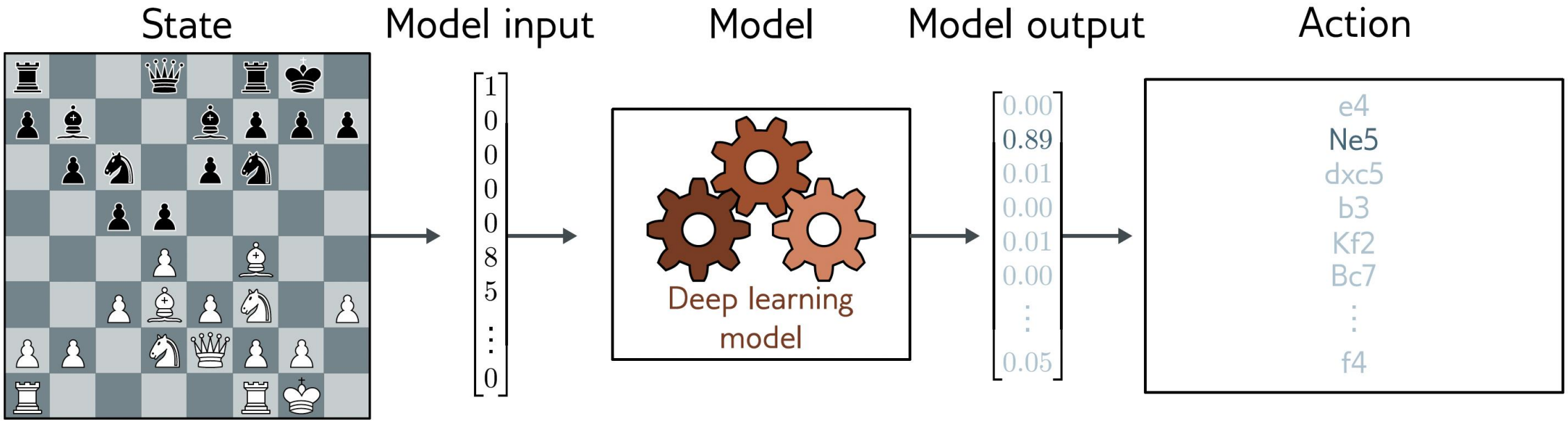
:

Action

e4
Ne5
dxc5
b3
Kf2
Bc7
⋮
f4

例子：国际象棋

- 状态是棋盘上棋子的当前布局
- 动作是当前符合规则的走棋方式
- 成功吃掉对方棋子可以得到正奖励，而自己的棋子被对方吃掉会得到负奖励



上机实验部分

- 使用Python作为主要编程软件
- 推荐VSCode代码编辑器 + GitHub Copilot插件进行AI辅助编程
- 需要注册 GitHub 账号
  - 校园网络可访问GitHub（ https://github.com ）
  - 用学校邮箱注册为学生（https://github.com/education/students）可获取Copilot无限使用权