

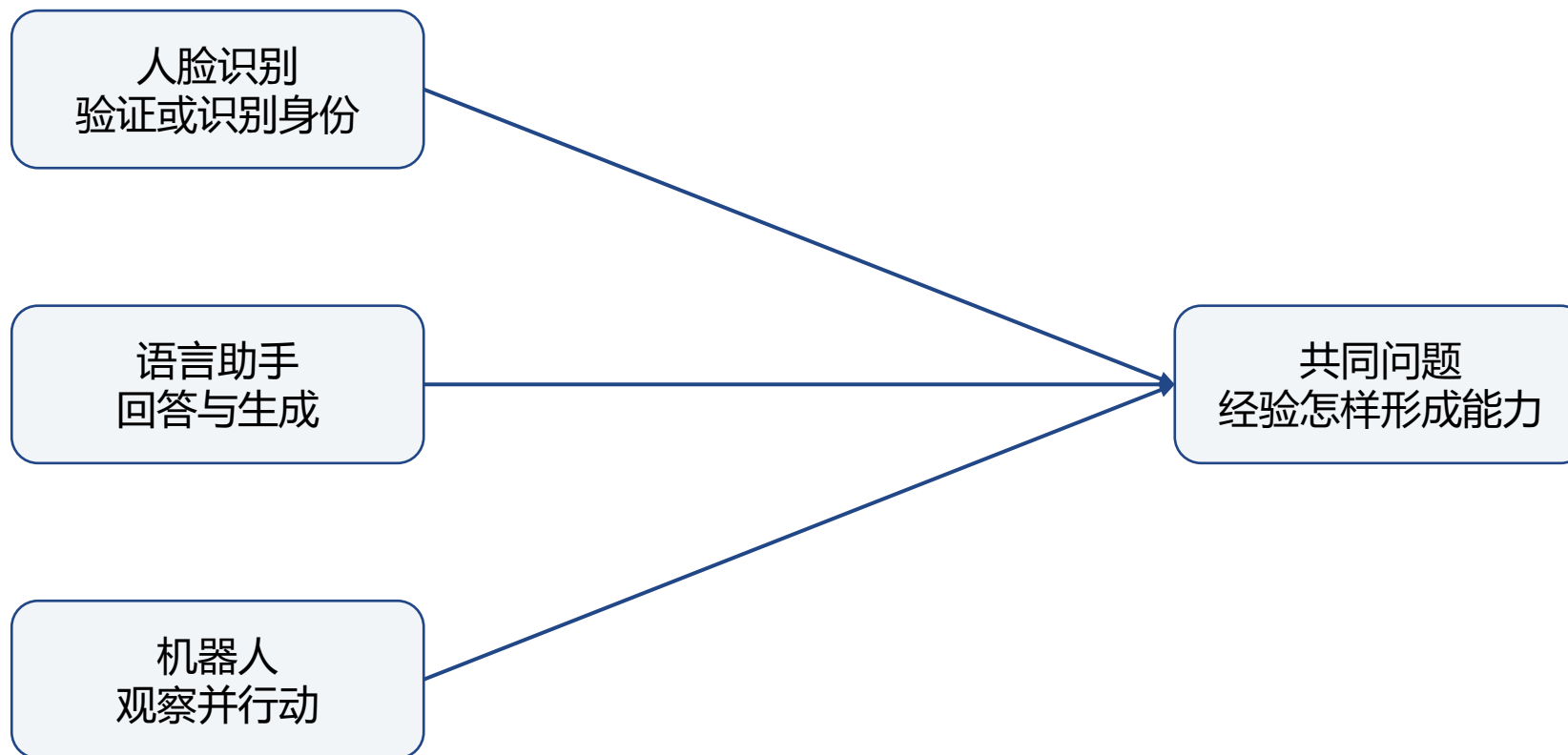
01 机器学习和统计学习

从现实能力到机器学习的基本过程

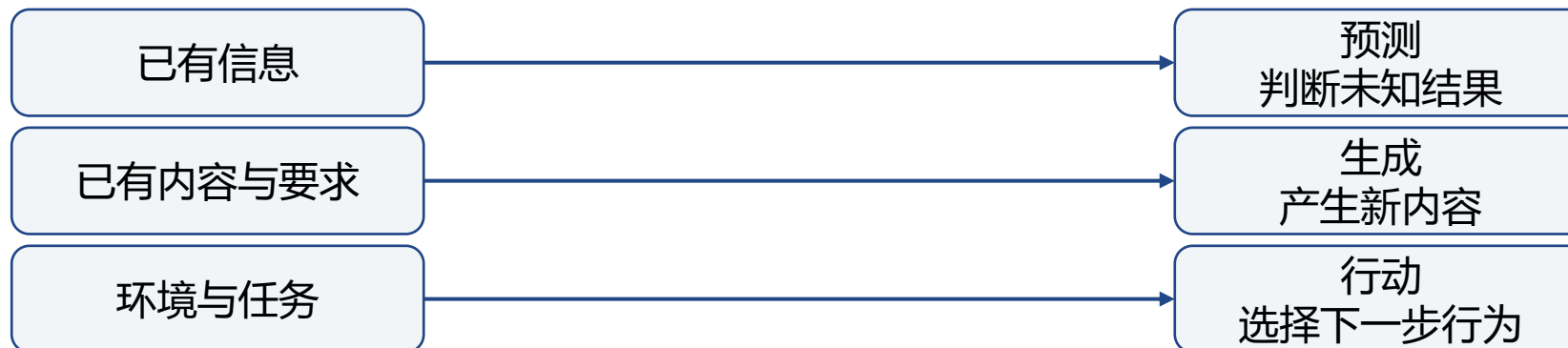
1 认识机器学习

现实能力与共同问题

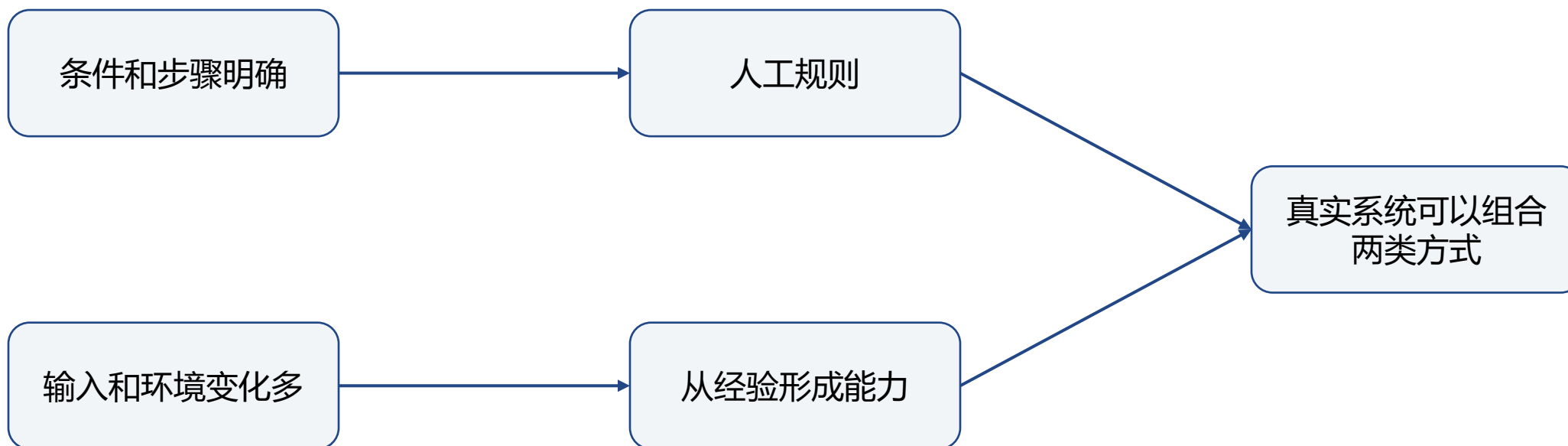
- 我们已经能够直接观察到机器学习系统表现出的多种能力
- 三类应用共同引出一个问题：系统怎样从数据或交互经验中形成可复用的处理能力



- **预测**：根据已经获得的信息，对类别、数值或未知结果作出判断
 - 如判断两张人脸是否属于同一人
- **生成**：根据已有内容和要求，产生新的文本、图像、声音或其他内容
 - 如根据问题生成一段回答
- **行动**：根据当前环境和任务选择下一步行为
 - 如机器人观察桌面后移动机械臂抓取目标物体
- 预测、生成和行动可以同时出现在一个系统中

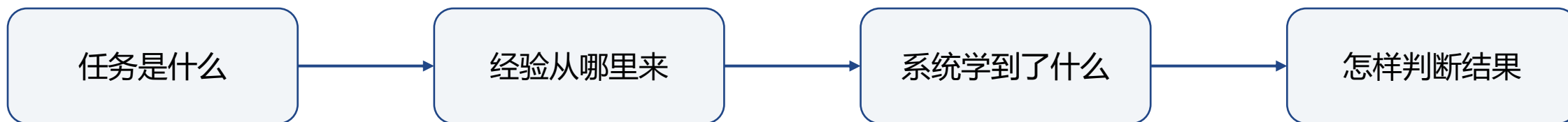


- **规则可以明确表述**：人可以预先写出条件和步骤，如计算税额或检查固定格式
- **变化情形难以穷举**：人说明任务并提供数据或交互机会，让系统依据经验调整处理方式
- **真实系统**：人工规则、搜索、规划和从经验形成的能力可以组合使用



三类系统的四个共同问题

- **任务是什么**：系统最终需要作出什么判断、生成什么内容或完成什么行动
- **经验从哪里来**：系统可以利用哪些样本、记录、反馈或交互过程
- **学到了什么**：经验怎样转化为以后可以继续使用的规律、处理方式或能力
- **怎样判断结果**：用哪些可观察现象或指标比较系统在任务和新情况中的表现



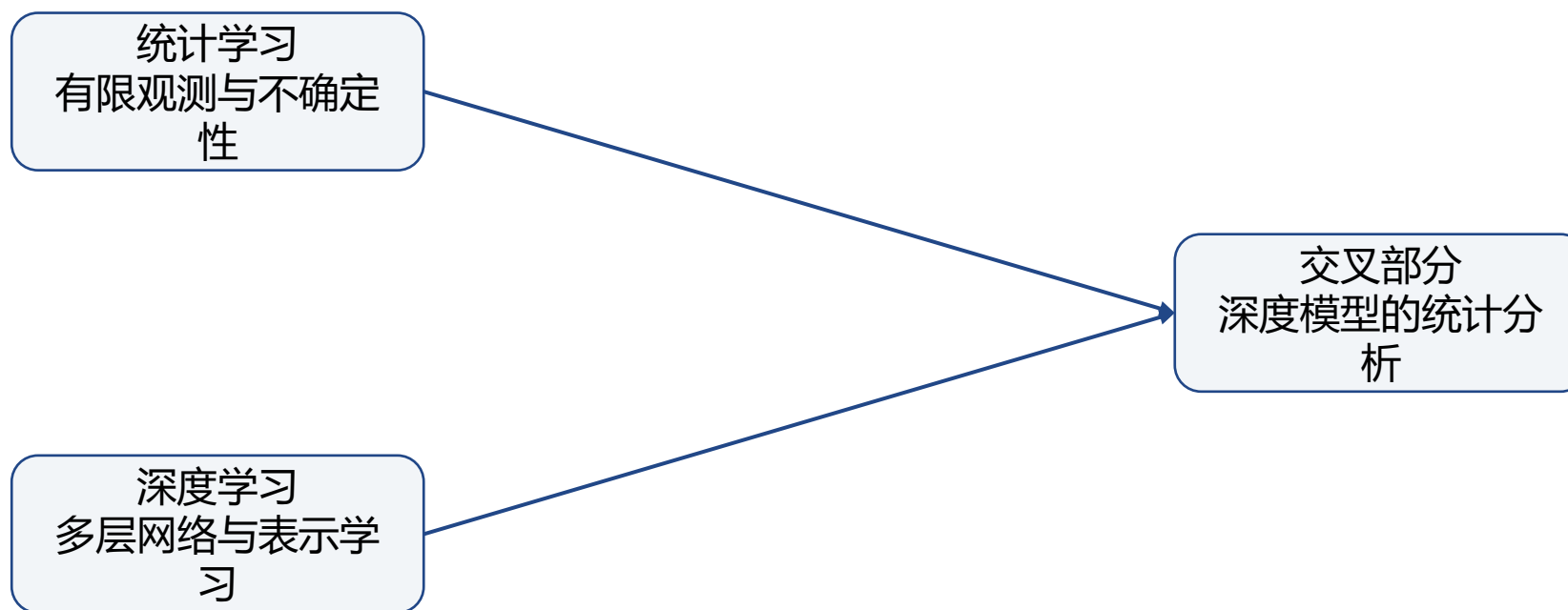
1.1 领域与技术路线

第一课先建立领域坐标与技术路线，再回到具体系统的共同问题

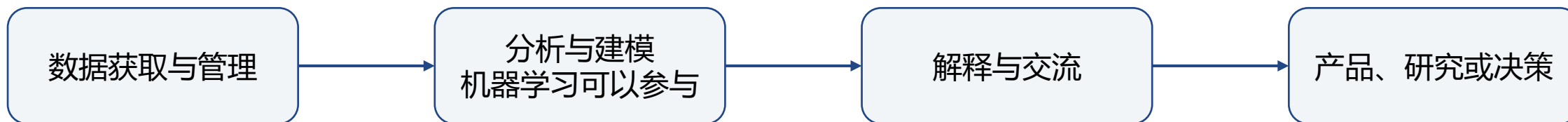
- **人工智能**关注系统的感知、推理、规划、学习、生成与行动能力
- **机器学习**利用数据或交互经验，形成或改进系统完成任务的能力
- **关系与边界**：机器学习是实现人工智能能力的重要路线，也用于科学分析和预测；人工智能系统还可以使用搜索、规划、知识表示与人工规则



- **统计学习**从有限观测中识别并利用数据关系，同时判断不确定性与适用范围
- **深度学习**以多层神经网络为基础，依据数据形成复杂表示并完成任务

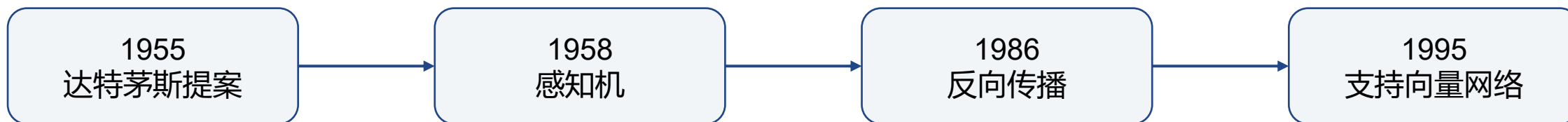


- **数据挖掘**关注从实际数据中发现有用的模式、关系或知识
- **数据科学**覆盖从数据获取、管理和处理到分析、解释、交流和应用的完整过程
- **机器学习的角色**：提供分类、预测、聚类和异常检测等方法；数据获取、管理、解释与应用仍需其他工作

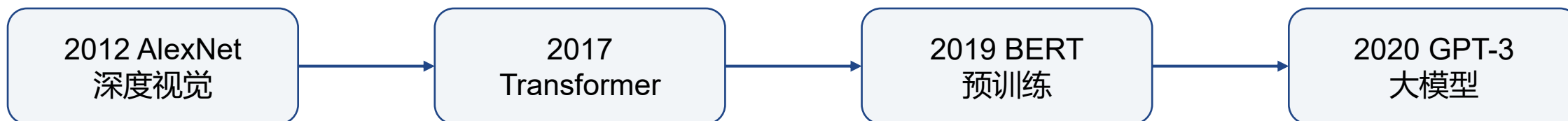


领域	主要关心的问题	常见结果或范围
人工智能	怎样构造具有多种智能能力的完整系统	感知、推理、学习、规划与行动
机器学习	怎样从数据或交互经验形成任务能力	预测、生成或决策模型
统计学习	怎样从有限观测发现关系并判断适用范围	预测、推断与结构分析
深度学习	怎样用多层计算从复杂数据形成任务能力	视觉、语言与语音系统
数据挖掘	怎样从实际数据中发现有用模式或知识	模式、规则、分组与异常
数据科学	怎样完成从数据获取到解释和应用的完整实践	分析、数据产品与决策支持

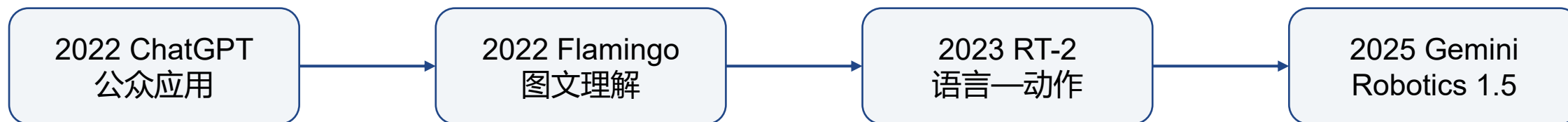
- 人工智能长期并用人工规则、搜索和机器学习；这里聚焦“机器怎样从样本中形成判断”
- 四个节点提供早期机器学习的时间坐标，不表示唯一继承关系



- 大规模数据与深度网络推动深度视觉发展，预训练方法随后扩展到语言模型；箭头仅表示年代先后



- 2020 年代，大模型进入公众应用，相关能力进一步扩展到图文理解和机器人行动
- 四个节点分别标记对话、多模态和机器人方向，不表示一条单一路线



- **数学证明**：AlphaProof (2024) ; Navier–Stokes 解答与 Lean 证明 (2026, 待审)
- **世界模型**：渲染器、模拟器和规划器；Genie 3 属于交互式渲染器
- **编码智能体**：vibe coding、Claude Code 与 Codex；结果仍需测试与审查

1.2 任务、数据与特征

已有领域坐标后，回到具体应用：系统做什么、凭什么学习、如何评价

- **任务**说明系统接收什么输入、需要什么输出
- 猫狗识别接收一幅动物图像，并从“猫”和“狗”中输出一个类别，因此属于**二分类任务**
- 类别答案在训练时提供学习信号，在评价时用于核对预测；使用模型时，答案不作为输入
- 若改为判断动物品种或定位动物，输入图像可以相同，但输出和任务已经改变



- **学习经验**是学习过程用于形成或调整表示、模型或策略的信息
- **带答案或无答案的数据**都可能提供经验
 - 带答案数据：猫狗图像及其类别标签
 - 无答案观测：利用相似性、共现关系或重复结构形成学习信号
- 经验还可以来自**数据内部构造的目标或行动反馈**
 - 如遮住部分内容后要求恢复，或机器人抓取后观察任务是否完成
- 经验所提供的信息不同，学习方式也会不同

- **性能指标**把任务表现转化为可比较的量，如错误率、准确率或处理时间
 - 对猫狗识别系统，可以分别统计猫被认成狗、狗被认成猫的情况
- 指标必须对应任务，不同错误在实际使用中的影响也可能不同
- 系统训练或调整时采用的目标量、评价时报告的指标和真实使用代价承担不同职责
 - 三者可以相同，也可以不同
- 判断系统是否有所改善，需要预先明确评价依据

- **任务** T : 系统需要完成什么工作
- **经验** E : 系统依据哪些数据或交互经历形成能力
- **性能指标** P : 怎样观察和比较系统在任务上的表现
- 本页的 P 来自英文 performance, 表示性能指标。第二部分的 $P(\cdot)$ 表示概率
- 以猫狗识别为例:
 - T 是根据一幅图像判断主要对象是猫还是狗
 - E 是带有猫狗类别答案的图像
 - P 可以是在约定测试条件下的分类准确率
- T 、 E 、 P 用于检查任务、经验和性能指标是否明确, 是本章使用的一种分析框架

- 现实中存在不同的动物品种、姿态、背景、设备和拍摄条件
- 一次照片、测量或交互记录只保存了现实对象在特定条件下的一次观测
- **数据集**是为某个任务组织起来的有限观测集合
- 数据是否具有代表性，需要相对于任务、采样方式和预计使用范围判断
 - 数据可能遗漏某些品种、姿态、光照、背景或拍摄设备
- 数据量与覆盖范围需要分别判断；当有限数据与目标使用范围具有适当对应关系时，仍可提供有效证据

- **原始输入**是系统直接接收的数据，如图像、声音、文本或传感器记录，而不是现实对象本身
- **特征或表示**是从原始输入获得、供后续处理使用的数据描述或内部表达
 - 可以通过选择、提取、变换或学习得到
- 特征或表示旨在突出当前任务需要的信息，但也可能保留无关因素或遗漏重要信息；同一个对象在不同任务中可以采用不同表示

- **人工特征**是由人依据任务知识设计、选择或变换得到的特征
 - 如动物轮廓、局部纹理、耳部形状或颜色统计
- **学习表示**是模型在训练过程中从数据中学习得到的内部表达
 - 如神经网络中间层产生的数值编码
- 两种方式都会受到数据、学习目标、模型结构或任务知识的影响，也都可能保留有用信息或丢失重要信息
- 传统机器学习通常更依赖人工特征；现代深度学习更强调特征与任务处理共同学习
- 特征或表示限定后续过程能够直接利用的信息，因此连接了数据、模型和任务

1.3 学习、模型与学习结果

系统怎样从有限经验形成能力，又用什么保存和使用所形成的能力

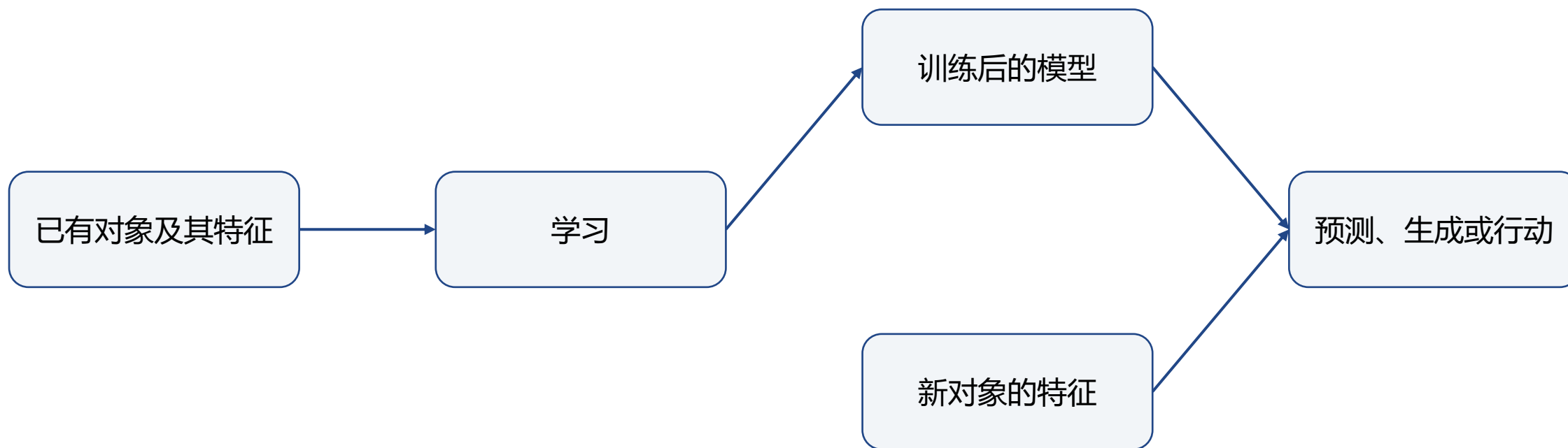
- **目标**：把经验用于未见但相关的情况，而不是复述已有样本
- **为什么可能**：某些与任务有关的关系会在相近条件下重复出现
 - 如猫或狗的类别通常不会因一次视角或光照变化而改变
- **为什么不能只靠数据**：有限数据不能唯一确定普遍正确的规律
- **方法还要做什么**：限制可表示、比较和选择的关系，为有限证据加入必要的假设或偏好
- **结论**：泛化以训练经验与新情况足够相关为前提

- **模型是什么：**模型是用数学或计算结构表示输入与输出的关系、数据规律或决策规则，并据此处理新对象的形式
- **常见形式：**
 - 公式及其系数
 - “如果.....那么.....” 的树形规则
 - 不同情况发生可能性的概率关系
 - 由大量相互连接计算组成的网络
- **怎样用于新对象：**模型把新对象的特征或表示映射为预测、生成结果或行动选择
- **与训练算法的区别：**模型承载并使用学习所得的结构、关系或决策规则；训练算法负责求得具体模型

层次	回答的问题	猫狗识别例子
平实描述	系统学会了什么	判断图像中的主要对象是猫还是狗
数学表示	学习对象怎样形式化	x 表示图像或特征, $f(x)$ 输出类别或分数
实现形式	结果怎样存储并参与计算	公式系数、树规则、训练样本或网络权重
可观察行为	学习结果起什么作用	对新图像输出类别和分数

从经验到新对象：特征、学习与模型

- 特征或表示说明系统使用哪些信息描述和处理对象
- 学习依据已有经验形成模型；模型承载学习所得的关系，并用于处理与训练经验足够相关的新对象



1.4 学习范式、任务与建模路径

按经验来源、任务类型和建模路径理解不同名称

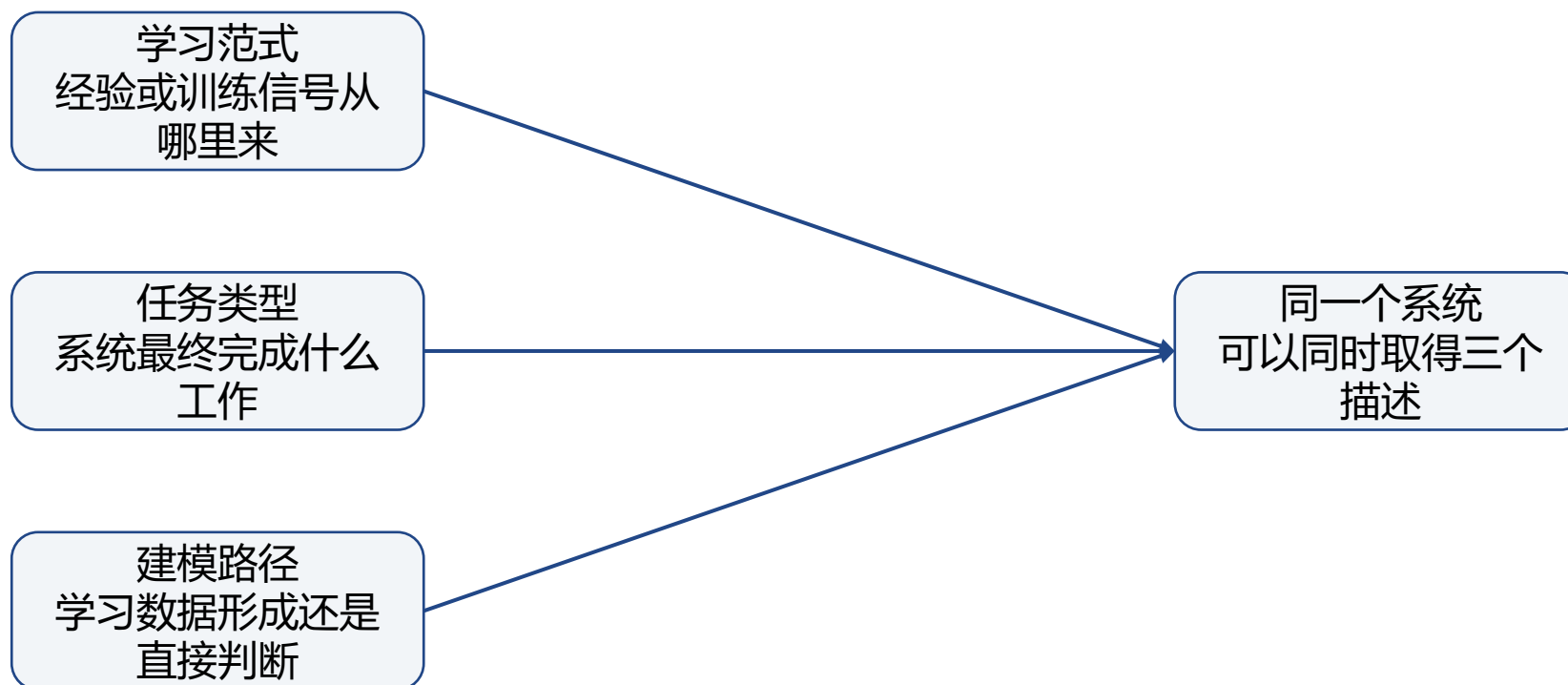
- **监督学习**从输入及其对应答案组成的样本中学习预测关系
- **无监督学习**从没有任务答案的观测中寻找结构、分组或数据规律
- **自监督学习**利用数据内部关系构造训练答案，如根据上下文预测被遮住的内容
- **半监督学习**联合利用少量带答案数据和大量无答案数据完成任务
- **强化学习**通过行动及其反馈，学习在环境中怎样选择行为
- 可以从学习信号或经验来源的角度理解监督、无监督、自监督、半监督和强化学习；
这些名称不直接说明系统最终完成什么任务

- **分类**：从有限类别中作出选择，如判断图像中的主要对象是猫还是狗
- **回归**：预测连续数值，如预测房屋价格
- **聚类**：把相似对象组织成组，事先不指定每组的类别答案
- **生成**：产生新的文本、图像、声音或其他样本
- **决策**：根据当前信息选择行动
- **表示学习**通常是中间学习目标，为其他任务形成更适用的内部表达
- 任务与学习范式是两条分类轴：同一任务可采用不同范式，同一范式也可服务多种任务

- 在监督分类中，**生成方法**学习数据或各类别数据怎样产生
 - 如分别描述猫图像与狗图像通常具有怎样的特征，再根据各类特征分布判断输入类别
- **判别方法**直接学习怎样从输入得到类别或判断结果
 - 如直接学习一条把猫与狗分开的判断边界
- 生成方法可以支持生成、缺失信息处理或分类，判别方法通常更直接地服务预测
- 两条路径所需的假设和信息不同，优劣取决于具体任务与数据条件
- 现代生成模型学习复杂数据分布并产生新内容，范围比传统监督分类中的生成—判别对照更广

- 智能体根据当前信息选择行动；与决策有关的信息称为**状态**，可选行为称为**动作**
- 环境用**奖励**反馈当前结果是否符合目标；奖励不一定等于长期效果
- 依据状态选择动作的规则称为**决策策略**
- 动作会改变环境，也会改变系统下一步获得的数据
- 具身智能还涉及感知、控制、规划以及身体与环境交互；强化学习只是可选方法之一

- 三条轴回答不同问题，可以同时描述同一个系统
- 猫狗识别可以采用监督学习完成分类，并沿判别路径直接学习类别判断
- 监督学习、分类和生成路径分别属于不同维度；深度学习则描述所采用的方法体系



2 机器学习的基本过程

先看训练与使用全景，再进入概率、模型、训练和评价

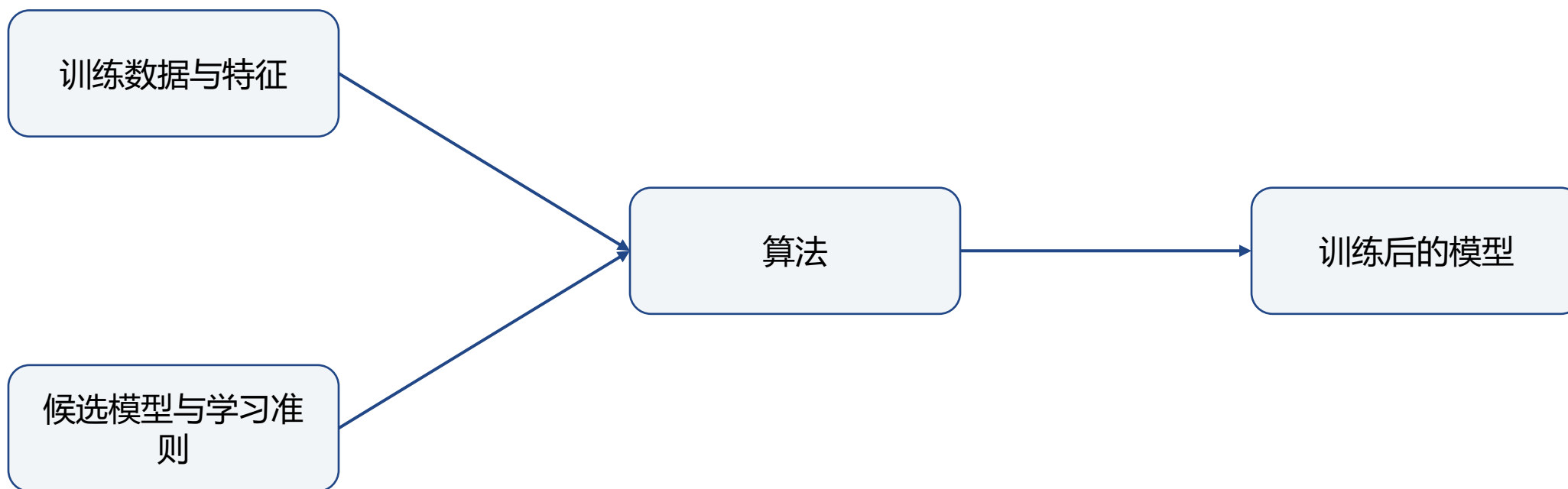
2.1 训练、概率与模型

模型怎样由有限数据、学习目标和算法共同确定

- **训练阶段**使用已有的猫狗图像及其类别答案，确定模型中的具体关系或内部数值
 - 训练需要数据及其特征或表示、候选模型形式和学习目标
 - 输出是一个训练后的猫狗分类模型
- **使用阶段**把训练后的模型用于未参与训练的新图像
 - 输入是新图像及其特征或表示
 - 输出是类别、分数或概率
- 在本章采用的典型流程中，训练阶段形成模型，预测阶段使用训练完成的模型

训练结果的形成：数据、候选模型、学习准则与算法

- 有限数据通常与多个候选结果相容，不能单独给出唯一的训练结果
- 训练数据与特征、候选模型和学习准则共同限定选择；算法据此求得训练后的模型



- **有限数据集**包含已经观测到的动物图像和类别，是现实中可能出现的猫狗图像的一部分
- **数据来源或产生条件**描述图像在什么动物、场景、设备和收集方式下形成
- **目标数据条件**描述模型预计在哪些图像和使用条件中工作
- 数据是否具有代表性，取决于数据产生条件与目标条件的关系，包括重要对象、变化范围及其出现比例是否得到适当反映
- 训练数据与目标条件不必完全相同，但二者的关系必须足以支持预期使用

2.1.1 用概率描述有限数据与未知规律

要描述图像特征与类别怎样共同变化，先把它们记为变量 X 和 Y 。联合分布 $P(X,Y)$ 描述不同取值共同出现的概率，并为条件预测和生成建模提供共同起点

用 100 幅假设图像的计数建立联合分布直觉

教学特征 X (共 100 幅)	$Y = 1$: 猫	$Y = 0$: 狗
$X = 1$: 耳部较尖 (共 40)	30	10
$X = 0$: 其他 (共 60)	20	40
列合计 (共 100)	50	50

- 以下用 $\hat{P}$ 表示假设样本中的经验比例；总体概率 P 仍描述目标数据条件中的规律
- **边缘分布**回答 “一个变量总体怎样分布”
 - $\hat{P}(X = 1) = 40/100 = 0.40$
- **条件分布**回答 “已知一个变量后，另一个变量怎样分布”
 - $\hat{P}(Y = 1 | X = 1) = 30/40 = 0.75$ ；竖线 “|”读作 “在.....条件下”
- 条件方向决定问题与分母， $P(Y | X)$ 和 $P(X | Y)$ 一般不同

- 同一个联合事件可以从两个方向分解：
 - $P(X, Y) = P(Y | X)P(X) = P(X | Y)P(Y)$
 - 在教学样本的经验比例中： $0.30 = 0.75 \times 0.40 = 0.60 \times 0.50$
- 生成与判别路径可以用同一联合分布中的不同条件关系表示
 - $P(Y | X)$ 直接描述给定图像特征后的类别概率，对应判别路径
 - $P(X | Y)P(Y)$ 先描述各类数据怎样形成，再通过贝叶斯关系得到类别判断，对应生成路径
- 联合分布及其条件分解描述概率对象，不表示有限数据已经恢复真实规律
- 从样本比例推广到新图像，还需要关于采样、依赖关系和稳定性的假设

2.1.2 从候选模型到训练结果

概率关系描述了需要学习的关系，但不会自动给出可执行的预测器。训练还要限定候选模型，用学习准则比较候选，再由算法求得一个具体结果

- 训练先明确允许比较和优化的候选模型范围
- **模型族或假设空间**是训练可比较和选择的一组候选模型
 - 猫狗识别可以采用概率模型、线性模型、树模型或神经网络
- 参数化模型写作 f_{θ} , 参数 θ 区分模型族中的具体候选
- **参数**由当前训练过程确定; **超参数**在参数训练之外设定, 如正则化强度或树深限制
- 训练后的模型是候选集合中的一个具体结果, 不等于整个模型族

- 模型族给出了候选集合，还需要一个统一尺度来比较哪些候选更符合任务，因此引入损失与风险
- 对一次预测，**损失函数** $\ell(y, f(x))$ 衡量预测结果的代价
 - 把猫判成狗与把狗判成猫，可以设置相同或不同代价
- 对有限训练数据，**经验风险**汇总模型在所有训练样本上的平均损失
 - 设训练集 D 含 n 个样本， x_i 表示图像或其特征， y_i 表示类别， $f(x_i)$ 表示模型输出
 - $\hat{R}_D(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i))$
- 损失衡量一次预测的代价，经验风险概括模型在训练数据上的平均损失

- **期望风险**：目标条件下反复使用模型时的平均损失
 - $R(f) = \mathbb{E}[\ell(Y, f(X))]$, 其中 $\mathbb{E}$ 表示按目标数据条件的概率规律求平均
- 目标概率规律通常未知，只能借助有限数据估计
- 经验风险能否反映期望风险，取决于数据、模型和学习过程

- 损失和风险规定怎样比较预测结果；对于概率模型，学习准则还可以基于已观测数据的概率建立
- 假设 $\theta = P(X = 1 | Y = 1)$ ，表示猫图像满足该简化视觉特征的概率
- 用 D 表示已经观察到的 50 幅猫图像，其中 30 幅满足该特征、20 幅不满足
 - 每幅满足特征的图像贡献一个因子 θ ，每幅不满足的图像贡献一个因子 $1 - \theta$
 - 在假设每幅图像的观测相互独立时，50 个因子相乘得到：
 - $L(\theta; D) = \theta^{30}(1 - \theta)^{20}$
- 固定已经观察到的数据以后，**似然**把概率模型赋予已观测数据的概率或概率密度看作参数的函数
- 似然用于比较参数怎样解释固定数据，不是“参数本身的概率”

- **极大似然估计**选择使已观测数据似然最大的参数
 - 本例观察比例为 $30/50 = 0.60$ ，似然在 $\hat{\theta} = 0.60$ 处最大
- 取负对数可把最大化似然转为最小化损失
 - 对数保持正数的大小顺序，并把乘法变成加法
 - 在独立观测假设下，负对数似然可写成各样本损失之和
- 因此，极大似然给出了概率模型的一种学习准则，也把概率描述与损失最小化连接起来
- 贝叶斯估计还会引入参数的先验信息，本章不展开，详见阅读材料

- 确定候选形式和比较标准以后，还需要可执行的计算过程得到具体结果，因此要区分模型、学习准则与算法
- **模型**规定系统允许怎样表示和使用输入、输出或数据关系
 - 线性打分函数属于模型形式
- **学习准则**规定依据什么目标、约束和偏好，从候选模型中作出选择
 - 尽量减少训练图像分类错误属于一种学习准则
- **算法**规定怎样通过计算得到满足准则的模型或近似结果
 - 以减少训练图像分类错误为目标，反复调整模型数值属于求解算法
- 模型给出候选形式，学习准则给出选择标准，算法执行求解过程

- 有限训练数据可能使多个候选模型得到相近结果，学习系统仍需决定优先选择哪一个
- **归纳偏置**是面对相容候选时采用的偏好或假设，可来自模型形式、特征、相似性、训练方法或领域知识
- **正则化**通过目标、约束、模型空间或训练过程表达偏好，常用于降低过拟合风险，但不提供必然保证
- 模型族、学习准则、算法和归纳偏置共同限定训练结果；能否用于新数据仍需独立评价

2.2 使用、评价与泛化

训练数据已经参与模型形成，训练表现不能独立证明新数据表现；接下来需要独立评价并限定结论

- **训练集**：拟合模型参数；它参与模型形成，训练表现不能独立证明泛化
- **验证集**：选择模型与训练设置；反复据此选择会逐渐适应验证集
- **测试集**：方案冻结后估计新数据表现；所得结论仅适用于测试所覆盖的条件，且不得提前查看

- 训练目标或损失参与模型拟合，评价指标用于比较任务表现，真实使用代价反映错误造成的实际影响
- 三者可以相关，具体关系取决于任务、模型和评价设计
- 猫狗分类可以报告总体准确率，也可以分别统计猫被判成狗和狗被判成猫的情况
- 指标把表现压缩成可比较的数值，但平均数不能说明错误集中在哪些对象和条件

- 指标发现表现差异以后，还需要检查具体错误，才能判断问题来自哪些对象、数据条件或系统环节
- **错误分析**检查错误类型、数据子群和失败条件
- 对猫狗分类，可以继续比较：
 - 不同品种、姿态、背景和光照条件下的表现
 - 不同设备或数据来源中的错误变化
 - 两类错误是否集中在某些共同特征上
- 错误分析把总体指标连接到具体失败条件，也为过拟合和分布变化提供可检查的线索

- 如果独立评价显示模型在训练数据上表现很好、在相关新数据上表现明显下降，就需要判断模型是否过度适应了有限经验
- **示例：模型对背景纹理形成过强依赖**
 - 猫狗模型把训练图像中的背景纹理与类别建立了过强的对应关系
 - 模型在训练数据上可以得到很低错误，在背景不同的新图像上却可能分类失败
- **过拟合**是学习结果过度适应有限训练经验，并因而在相关未知数据上表现明显下降的现象
- 训练误差低本身不等于过拟合；只有与独立的相关新数据表现比较，才能观察到训练表现与新数据表现之间的差距
- 过拟合可能来自数据不足或噪声、模型表达能力过强、目标选择不合适以及反复选择等多种因素

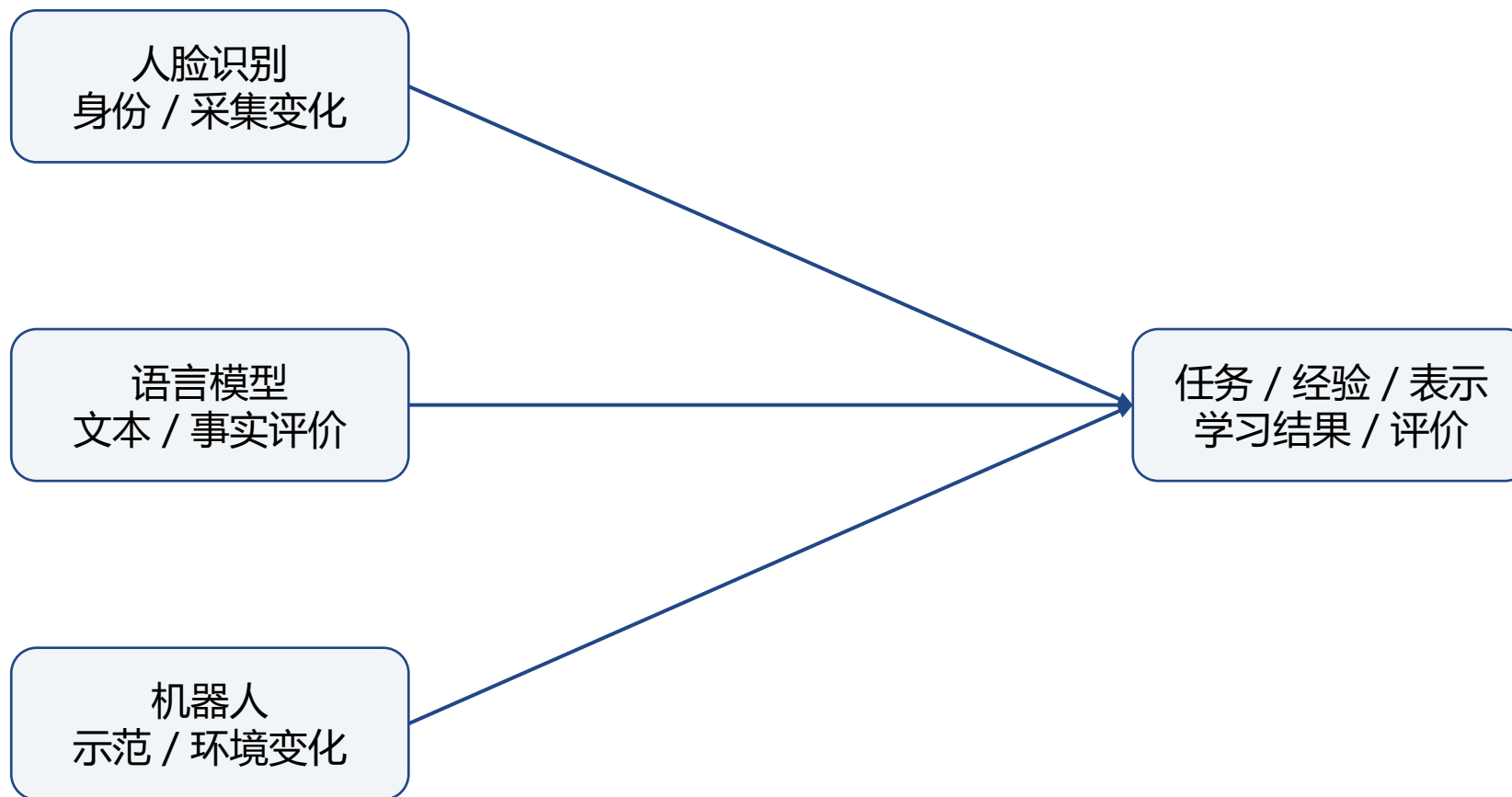
- 发现过拟合现象以后，需要进一步检查候选模型能够表达什么，以及训练时采用了什么选择偏好
- **模型容量**表示一类模型能够表达或区分多丰富的数据关系
- 容量太小可能无法表示任务所需关系；容量很大则可能同时表示有效规律和偶然细节
- 参数数量会影响部分模型的容量，但不能普遍完整衡量容量
- 正则化在训练时表达对候选结果的偏好，也会影响模型在新数据上的表现
- 合适的容量和正则化取决于任务、数据量、噪声、特征和评价方式
- 正则化可以帮助控制过拟合风险，但不能替代独立验证、测试和错误分析

- 即使模型通过一次受控测试，仍要追问这份证据能够支持哪些任务和数据条件
- **泛化能力**是模型在明确任务和数据条件下，对未参与训练的相关对象保持有效表现的能力
- 泛化依据来自训练数据、模型和学习方法，也来自独立验证与测试提供的有限证据
- 当使用阶段中品种、姿态、背景、光照或设备等条件的取值范围或出现比例相对训练阶段发生系统变化时，出现**分布变化**
- 受控测试只为测试所覆盖的相关环境提供证据
- 泛化讨论的是既定任务与相关数据条件；跨任务迁移和不受条件限制的“通用能力”属于更强的主张
- 使用条件变化时，需要重新检查数据覆盖、评价协议和模型表现

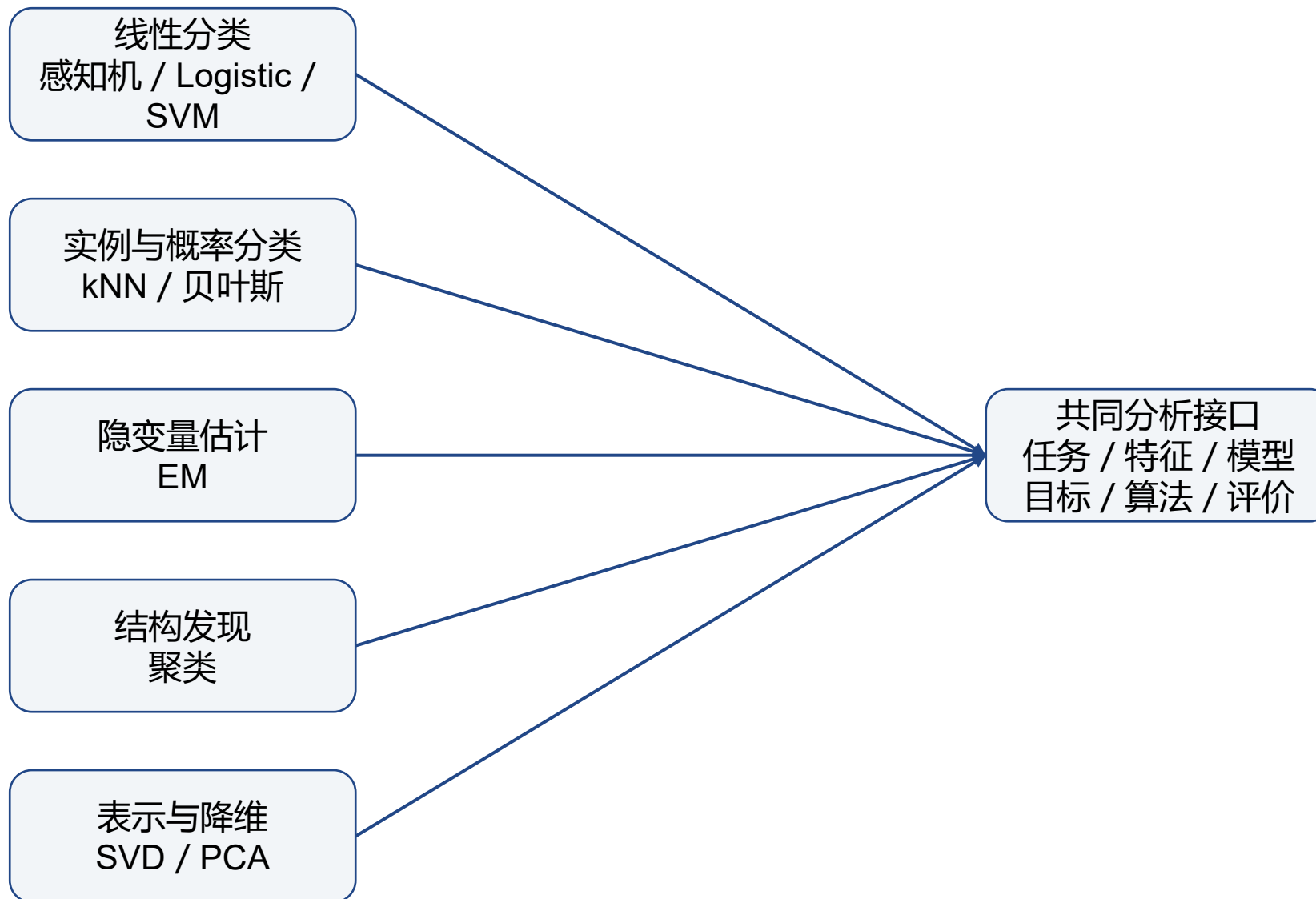
3 全章回收与后续学习

前两部分已经建立机器学习的共同语言和基本过程；最后回到开场案例，压缩分析框架并定位后续方法

- 回到开场案例：三类系统都可以从任务、经验、表示、学习结果和评价边界进行分析，但具体答案不同



后续方法围绕不同问题组织，共用同一分析接口



■ 任务是什么

- 系统最终需要完成预测、生成、结构发现还是行动

■ 经验从哪里来

- 有限数据、反馈或交互经验来自什么条件，准备支持什么使用范围
- 现实对象怎样形成输入，模型使用什么特征或表示

■ 系统学到了什么

- 有限经验不能唯一决定模型，还需要模型族、学习准则、算法和归纳偏置
- 学习结果采用什么模型形式，算法怎样得到并保存其中的关系或数值
- 训练后的结果怎样用于预测、生成、结构发现或行动

■ 怎样判断结果

- 训练表现不能独立证明泛化；数据分工、指标与错误分析提供评价证据
- 结论适用于哪些任务和数据条件，条件变化后是否仍然成立