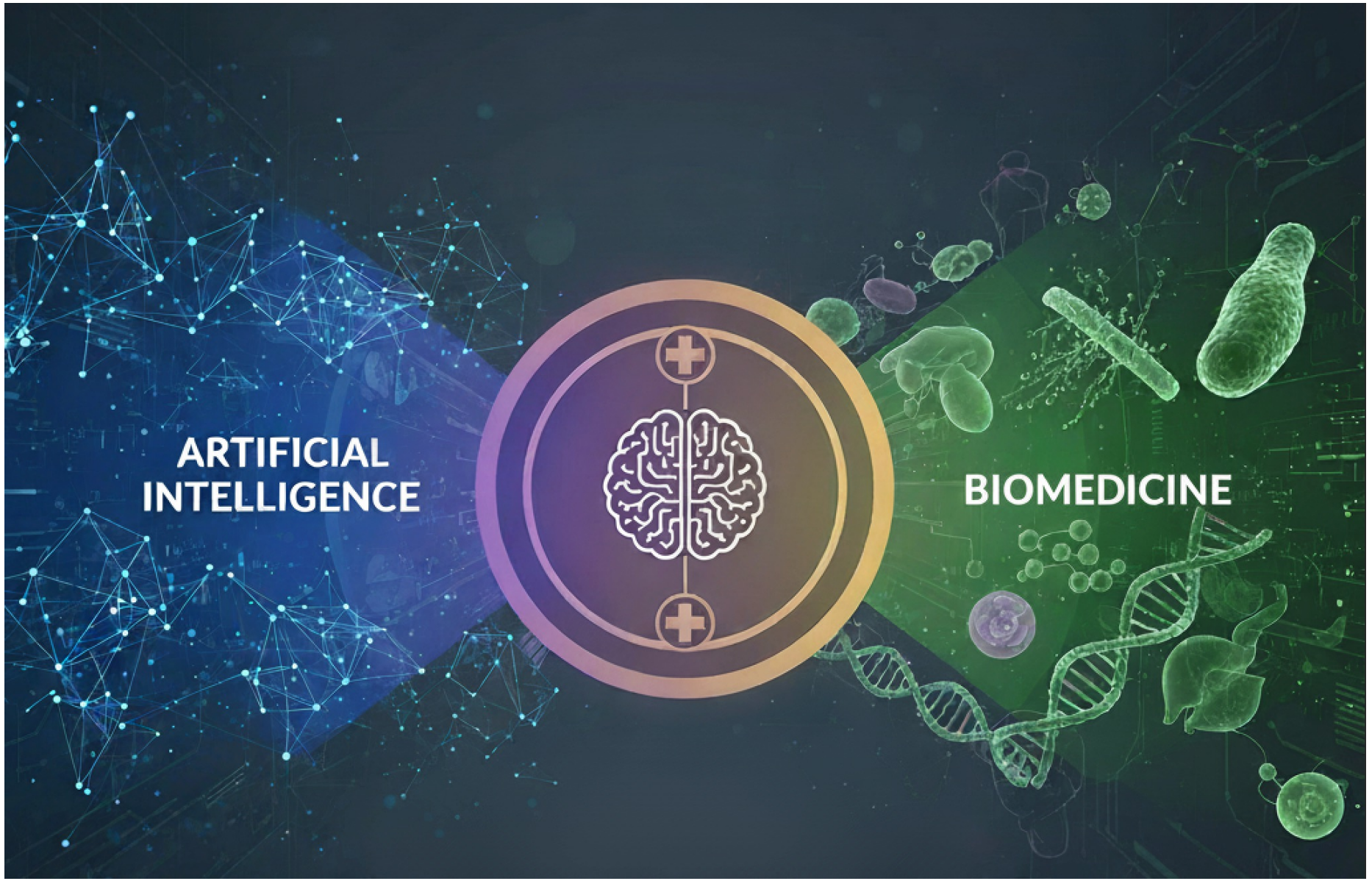


人工智能与生物医学的结合将具有深远意义

Generated by Google Gemini

第八课：医学人工智能中的可解释性

2025. 11

今日课程内容

- 什么是可信AI
- 对AI的预测结果进行解释
- 医学场景案例分析
- AI公平性定义
- 公平AI的方法框架
- 个体公平与群体公平

可信赖的机器学习

- 机器学习模型正越来越多地被部署到现实世界的应用中
  - 确保这些模型能够负责任地运行并值得信赖是至关重要的
- 需要开发和部署具备以下特性的 ML 模型和算法：
  - **准确 (accurate)**
  - **可解释 (explainable)**
  - **公平 (fair)**
  - **保护隐私 (privacy-preserving)**
  - **因果性 (causal)**
  - **稳健性 (robust)**

为什么理解模型至关重要

- 尤其是在涉及高风险决策的领域，理解模型如何做出决策非常重要
  - 医疗保健：为医生提供诊断与治疗建议，影响病人生命健康安全
  - 法律与司法：为法官提供裁决、量刑建议，影响公平正义、个人自由
  - 金融与信贷：提供贷款批准或拒绝决策，财务状况、经济机会

为什么必须理解机器学习模型

- 从黑盒到可信赖的决策伙伴
  - 提升模型“效用” (Utility):
    - 从技术和内部角度出发，确保模型本身是健壮、可靠和有效的
    - 关键点：可被调试、可检测偏见、可评估信任、可对部署进行审查
  - 满足“利益相关者” (Stakeholders) 的需求:
    - 从社会和外部角度出发，确保模型对人类是公平、透明和负责任的
    - 关键点：满足终端用户、决策者、监管机构、研究者的理解需求

为什么必须理解机器学习模型

- 提升模型“效用” (Utility):
  - **理解模型让我们可以调试模型**：当模型预测错误，如果不理解其工作原理，就无法修复它
    - **举例**：识别肺炎的X光片模型错误将个别健康图像诊断为肺炎，使用注意力热点可视化方法，发现模型在关注图像角落的医院标记而非病人的肺部，典型的“捷径学习”问题
  - **理解模型让我们可以检测模型的偏见**：确保模型对所有群体（如不同性别、种族、年龄）都是公平的
    - **举例**：癌症检测模型在测试集中表现良好，但是部署后发现只要是中国样本就检测为肝癌，理解模型后发现训练数据中，中国样本均患有肝癌，模型学到了该强关联，需要调整样本分布
  - **理解模型让我们可以评估模型的可信度**：即知道何时可以相信模型的输出，何时需要人工介入
    - **举例**：某疾病分类模型在某个病种下进行亚型分类时总分错，通过查看模型不确定性得分发现该疾病下置信度非常低，需要单独处理该疾病的训练和推理
  - **理解模型让我们可以开展部署审查**：在模型造成实际影响之前，全面评估其是否适合部署到现实应用
    - **举例**：...

为什么必须理解机器学习模型

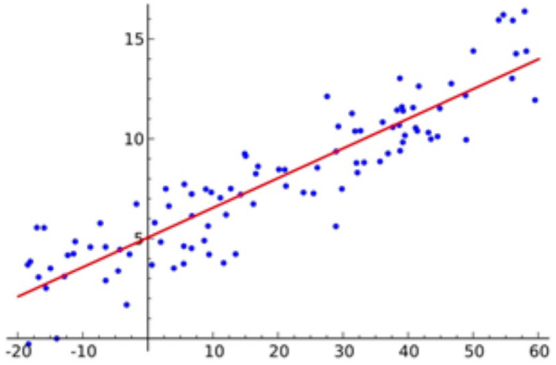
- 满足“利益相关者” (Stakeholders) 的需求:
  - **满足终端用户的需求**：他们需要知道“为什么这样”，以及如何“进一步处理”
    - **举例**：根据用户的基因检测报告AI给出了高遗传病风险的判断，需要同时告诉用户什么基因的异常和该基因与什么疾病关联，患病概率多高，生育计划建议等（理由+途径）
  - **满足决策者的需求**：他们是使用AI建议的最终责任人，必须能审查和否决AI
    - **举例**：AI建议对一名癌症患者使用某基因疗法，医生需要看懂AI的推理依据（如突变位点、相关连疾病知识，人群分布等），以判断这个建议是否优于标准疗法，并对患者负责
  - **满足监管机构的需求**：他们需要验证模型是否安全、合规、公平，才能批准其上市
    - **举例**：欧盟的《AI法案》要求对高风险AI系统（如医疗设备）进行严格审计，若一家公司不能清晰地解释其AI模型是如何工作的，该模型就无法通过合规审查，不能合法使用
  - **满足研究者与工程师需求**：他们需要理解现有模型的优缺点，以推动技术迭代和创新
    - **举例**：幻觉现象与神经网络中激活神经元分布的关联关系

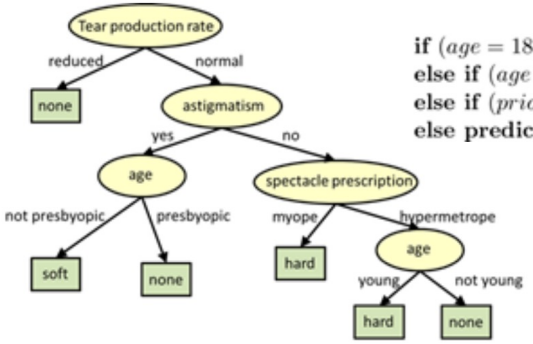
为什么必须理解机器学习模型

- 从黑盒到可信赖的决策伙伴
  - **理解机器学习模型，是为了确保技术始终服务于人类的最佳利益**
  - **对于技术（效用）：**
    - 理解 = 更健壮、更可靠、更公平的系统
    - 让我们能**修复错误**、**消除偏见**、**确定信任**
  - **对于社会（利益相关者）：**
    - 理解 = 透明、问责和信任
    - 让我们能**保护用户**、**赋能决策**、**满足合规**

如何实现对模型的理解

- **需要构建本质上可解释的预测模型**
  - 让人类能够轻松理解模型内容工作机制
    - 线性回归（直接看到参数）
    - 决策树（可追溯过程）
    - 决策规则（明确的路径）





if $(age = 18-20)$ and $(sex = male)$ then predict yes
else if $(age = 21-23)$ and $(priors = 2-3)$ then predict yes
else if $(priors > 3)$ then predict yes
else predict no

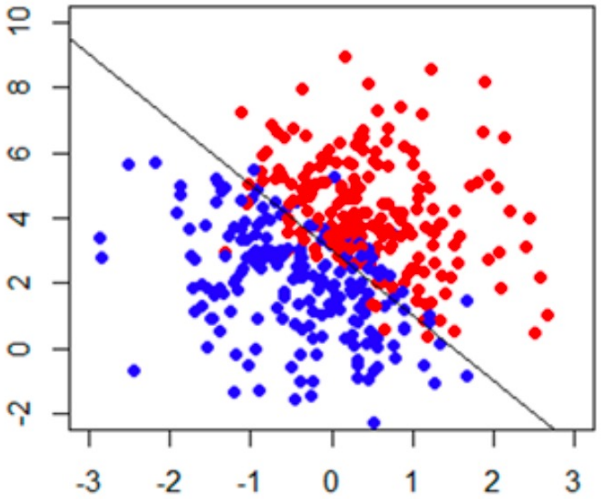
Decision rules

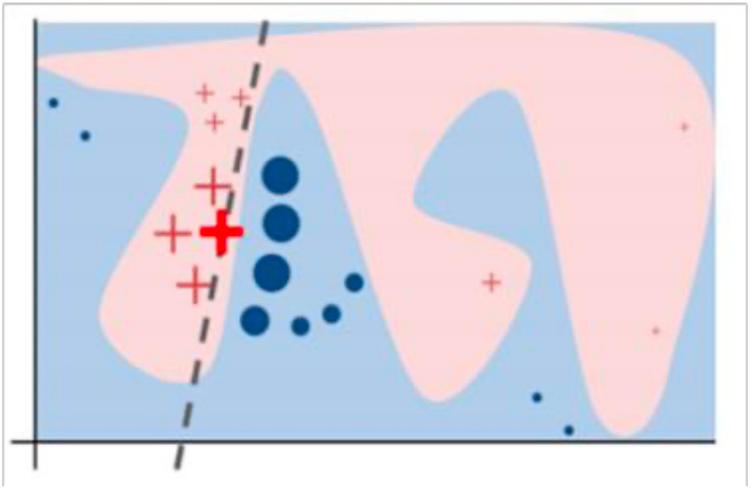
- **对黑盒模型的事后解释**
  - 显著性图（Saliency map）
    - 只能说明模型在看哪里，不能说明模型为什么这样判断
    - 相同的注意区域，不同的结果

高风险场景应该尽可能使用可解释的模型，而非去尽力解释黑盒模型

本质上可解释的模型 v.s. 去解释复杂模型

- **构建可解释且准确的模型**
  - 决策规则简单、简单公示可描述、理解模型判断依据，但性能需做出妥协
- **构建性能更好的复杂模型**
  - 虽然能力更强、更灵活，但判断完全黑盒，无法用语言解释，不知道为什么这样，容易过拟合、学到噪声、不可泛化





应该选择一个高准确率的“黑盒”，然后费力地去“解释”它，还是应该从一开始就选择一个“本质上可解释的”模型，即使其准确率可能没那么完美？

实现对模型的理解

- **可解释性与准确性的权衡以及黑盒模型的发展与流行，迫使我们依赖对机器学习模型进行“事后”（post-hoc）的“解释”**
  - **黑盒模型的流行：**由于研究场景的复杂化（多模态、大数据）和对高准确率的追求，深度神经网络模型已变得非常普及
  - **事后解释：**模型已经训练好，需要作为一个外部观察者，通过给它不同的输入，来观察它的输出，从而反向推断或近似模拟它的决策逻辑
    - “事后解释”并不是最理想的第一选择，但是一个必要的妥协
    - “事后解释”并不等于理解模型真正的内部工作原理，而是观察模型的行为是否合理

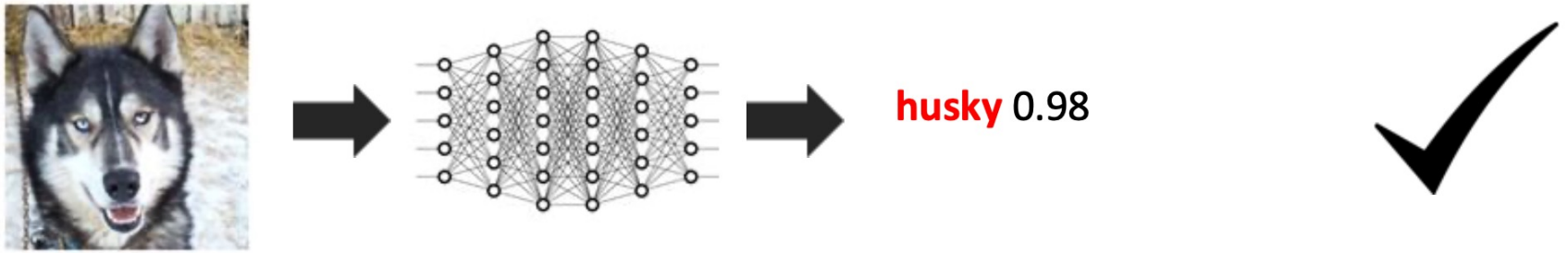
今日课程内容

- ✓ 什么是可信AI
- 👉 对AI的预测结果进行解释
- 医学场景案例分析
- AI公平性定义
- 公平AI的方法框架
- 个体公平与群体公平

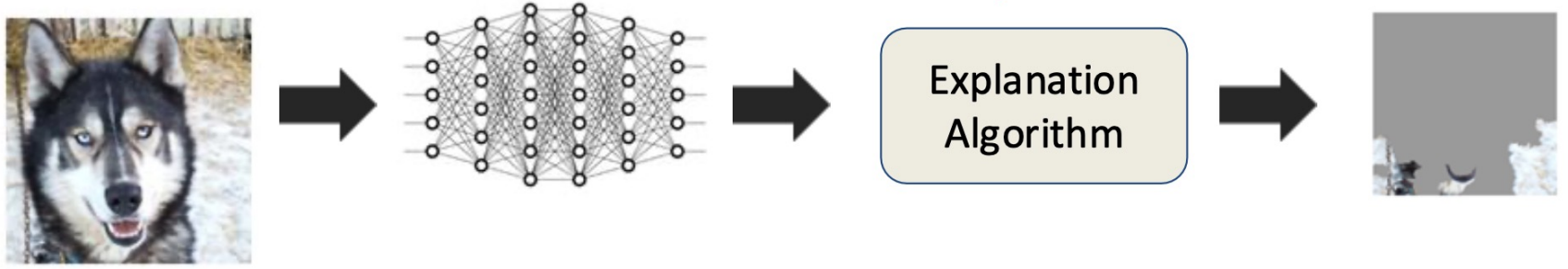
可解释的AI模型

- **可解释AI指一系列可以提供针对复杂模型行为的可理解描述的方法**

黑盒子模型，虽然得到了正确的答案，但不知道AI是如何做出这个决定的（根据狗的眼睛、毛发、还是耳朵来判断？）



通过算法生成一张新图，高亮显示原始图片中对AI决策影响最大的区域（背景与狗错误关联，换个背景会预测错误）



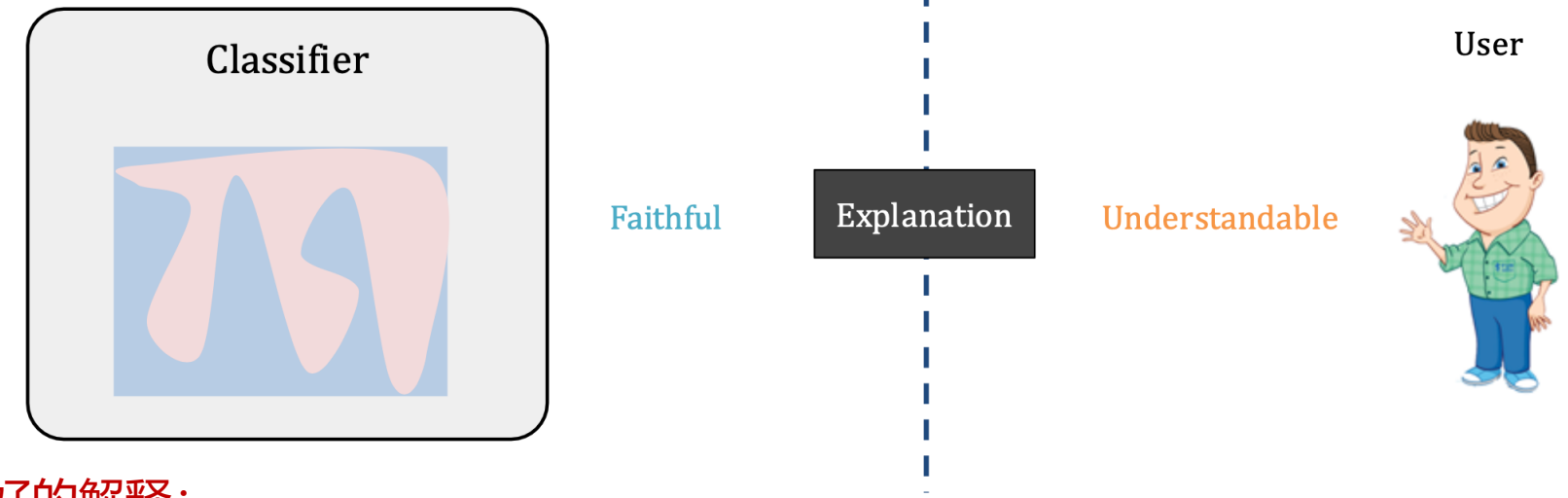
什么是可解释

- 对模型行为给出可被理解的描述

AI模型，人类难以直接看懂决策逻辑

连接模型和用户的关键桥梁

需要理解AI决策的人，如开发者、监管人员或最终用户



好的解释：

- 忠实性：解释必须准确地反映模型实际上是如何工作的，而非编造一个“听起来合理” 的理由
- 可理解性：解释必须以简洁、清晰、符合人类认知的方式呈现

可解释性方法的类型

- **局部解释（关注个体）**：解释模型为什么会对某一个特定的输入做出某一个特定的预测
  - 目标： 解释单个预测（为什么对这一个案例做出这样的预测？）
  - 作用1： 帮助发现在给定实例的局部邻域中存在的偏见（因为模型看到了雪，所以识别为哈士奇）
  - 作用2： 帮助审查单个预测是否基于正确的原因做出

- **全局解释（关注整体）**：解释模型作为一个系统的完整行为和总体逻辑
  - 目标： 解释模型的完整行为（做出决策的总体规则是什么？）
  - 作用1： 揭示影响更广泛子群体的宏观偏见（是否对某一个种族人群的检测持有偏见？）
  - 作用2：帮助从宏观层面 评估，这个模型是否适合部署

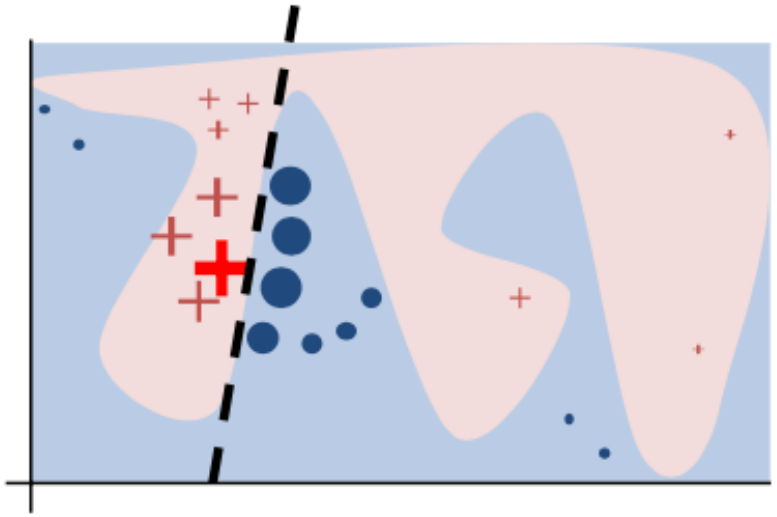
可解释性方法的具体技术

- **局部解释（关注个体）**：
  - 特征重要性评分：给输入的每个特征（像素、列）打分，告诉用户哪个特征对这一次的决策影响最大
  - 积分梯度：具体计算“特征重要性”的技术，通过数学方法将模型的预测结果归因到输入特征上
  - 原型解释：通过从训练数据中找一个“典型”或“原型”的例子来解释
  - 反事实解释：告诉用户需要对输入做什么改变，就会得到一个不同的预测结果

- **全局解释（关注整体）**：
  - 局部解释的集合：对数据集中的样本都运行一遍“局部解释”（如特征重要性评分），然后把这些解释汇总起来，寻找全局的模式
  - 基于表示的解释：通过查看神经网络中间层是如何对数据进行分类的，来理解模型内部是如何思考的
  - 模型蒸馏：训练一个全新的、更小、更简单的模型，让它去模仿那个复杂黑盒模型的行为

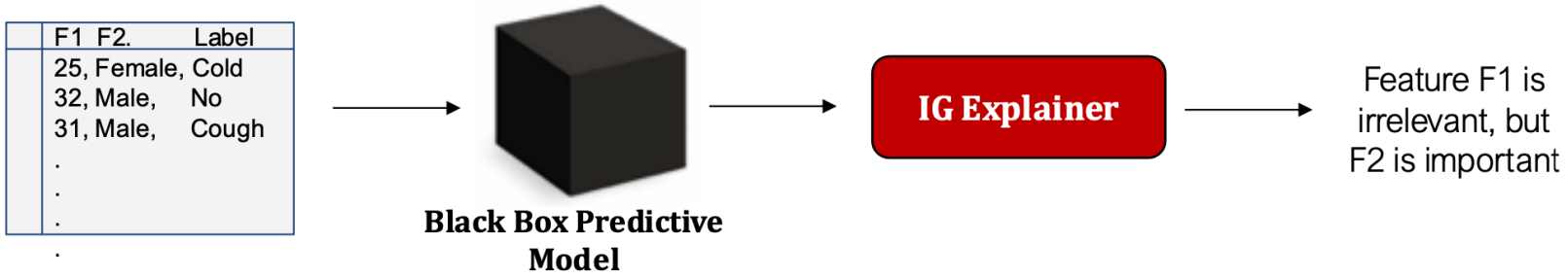
可解释性方法的具体技术

- **LIME：局部可理解的模型无关解释方法**
  - **思想：**虽然看不懂黑盒模型，但可以用做出决策的那个点（x_i）相似的数据，训练一个能看懂的简单模型（比如线性模型）来模仿它的行为
  - **方法：**
    - 决策点周围采样或生成相似数据点
    - 把这些数据喂给模型得到预测结果
    - 与决策点距离越近，权重更大
    - 用新样本、新预测结果、权重训练一个简单模型，模仿原模型在该局部区域的行为
    - 用简单模型来解释原模型



可解释性方法的具体技术

- **Integrated Gradients (IG)**：一种用于深度神经网络的模型解释方法，识别对模型预测贡献最大的重要特征
  - **原理**：通过计算从一个“基准线”（通常是零输入或空白输入）到实际输入路径上梯度的积分，来确定每个特征的贡献，从而解决了简单梯度方法在深度网络中可能出现的饱和问题
  - **IG 解释器**：分析输入数据和模型对该输入的响应（梯度），来计算每个特征的贡献度
  - **输出**：每个特征的归因分数，例如：“特征 F1 不重要，但 F2 很重要”
  - **优势**：
    - 可以应用于任何可微分模型，深度学习模型通常都满足可微分的条件
    - 将 IG 应用于已经训练好的、复杂的模型，不需要对原始的机器学习模型进行任何修改



可解释性方法的具体技术

- **Integrated Gradients有效的原因：**
  - **对特征贡献敏感：**如果改变一个特征导致模型预测结果发生变化，那么 IG 必须给这个特征分配贡献度（非零归因），通过插值路径积分，IG 捕捉了从“无”（基准）到“有”（实际输入）的完整影响过程
    - **插值路径：**建立一个从这个基准实例（如全黑图像或全零数据）到实际要解释的实例之间的插值序列，IG 沿着这个序列计算并累积梯度
  - **实现不变性：**解释的结果只取决于模型学到的功能（即它对输入做出什么预测），而不应该取决于模型是如何被编码或实现的，这保证了 IG 解释结果的公正性和鲁棒性
    - **功能等效模型：**如果两个模型在所有输入下的输出都相同，即使它们的内部实现非常不同，它们也是“功能等效”的
    - **归因相同：**IG 要求这两个功能等效的模型必须为相同的输入和基准产生完全相同的特征归因

可解释性方法的具体技术

- **Integrated Gradients的计算与可视化：**
  - **场景：**一个用于图像分类的机器学习模型
  - **目标：** 使用 IG 来解释模型为什么会给出特定的图像分类结果
  - **步骤1: 建立基准图像**
    - 计算需要从一个基准 (Baseline) 图像开始，该图像通常是中性的，其预测结果通常为0
      - 基准图像如全黑图、全白图、随机图等
    - 基准的目的：代表了模型在没有接收到有效信息时的“默认”状态，进而衡量处输入图像的像素值从这个中性基准移动到实际像素值时，对预测结果带来的累积变化

可解释性方法的具体技术

- **Integrated Gradients的计算与可视化：**
  - **场景：**一个用于图像分类的机器学习模型
  - **目标：** 使用 IG 来解释模型为什么会给出特定的图像分类结果
  - **步骤2: 生成线性插值**
    - 在基准图像（全黑图像）和原始图像（待预测图片）之间，生成一个线性插值序列
    - 插值图像：在特征空间中代表从基准到输入的“小步长”中间状态图像，如亮度递增过程

可解释性方法的具体技术

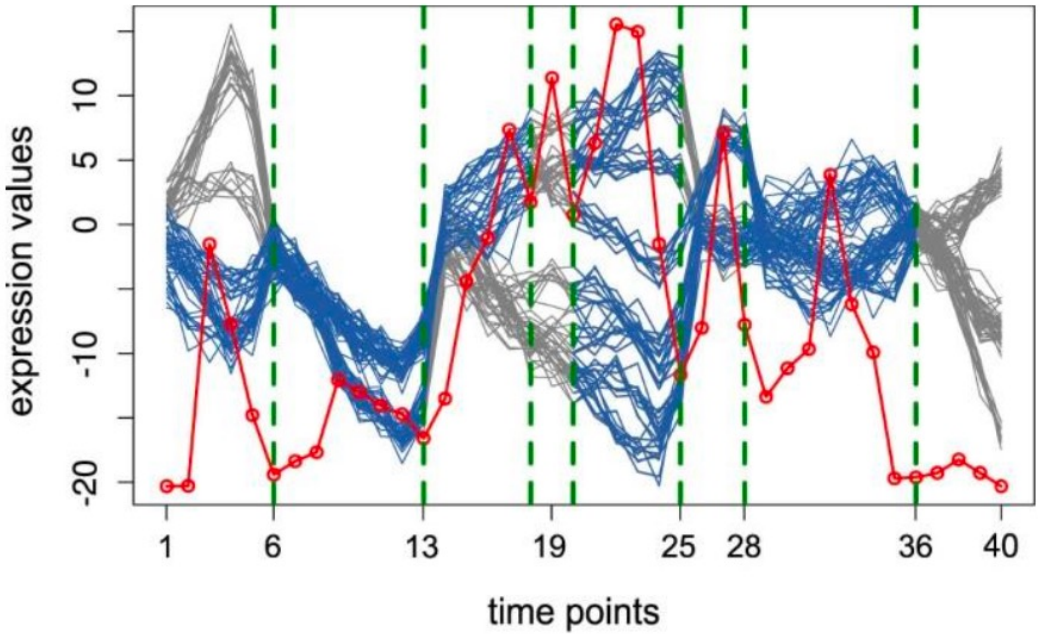
- **Integrated Gradients的计算与可视化：**
  - **场景：**一个用于图像分类的机器学习模型
  - **目标：** 使用 IG 来解释模型为什么会给出特定的图像分类结果
  - **步骤3: 计算梯度（将模型的黑盒特性转化为可量化的贡献度）**
    - 计算梯度以衡量特征的变化与模型预测的变化之间的关系
      - 梯度：$\frac{\partial(F(x))}{\partial x}$，代表了模型输出相对于输入特征的敏感度，即斜率
    - 告诉我们哪个像素对模型预测的类别概率具有最强的影响（导致模型输出变化的特征会获得归因值）
  - **步骤 4：计算数值近似（积分）（将所有局部梯度累积起来，得到最终贡献度分数）**
    - 通过求和或平均梯度来计算数值近似（获得整体影响大小）

可解释性方法的具体技术

- **基于原型的可解释性方法**
  - **思想：**通过展示与被解释实例相似的样本（可以是合成的或自然存在的），来证明模型做出决策的合理性
    - 如：“它看起来与训练集中的这三张图片（原型）非常相似，而它们都是猫，所以这张也是猫”
  - **技术方法：**
    - **影响函数（Koh & Liang 2017）：**识别训练集中的实例，这些实例对给定测试实例的预测结果负有责任（即影响最大），从而定位最关键的训练数据点
      - 如：模型错误地将一只狗识别为狼，影响函数可以找出训练集中导致这个错误分类的最有影响力的几张“狼”的图片，这对于数据调试和发现训练集中的偏见非常有价值
    - **激活最大化（Erhan et al. 2009）：**识别或合成能够强烈激活某个特定功能（如神经网络中的一个神经元）的示例（合成或自然的），旨在理解模型的内部工作机制
      - 如：一个神经元负责识别图像中的“眼睛”特征，激活最大化会生成一张能让这个“眼睛神经元”达到最大激活状态的图像，进而展示出该神经元学到了什么特定的视觉概念

可解释性方法的具体技术

- **基于原型的方法应用于时间序列模型的解释**
  - **挑战：**时间序列数据（如心电图）与图像和文本等不同，具有时间依赖性，这使得对其进行解释更加复杂
    - **时序数据本身具有内在复杂性：**
      - 噪声样本：包含大量的随机波动和噪声
      - 密集的信息特征：每个时间点都包含重要信息，且信息是密集连续的，难以隔离和判断
    - **时间模式不易被发现：**
      - 有意义模式只在时间段和长期行为中出现（如周期性、趋势性）
    - **扰动影响：**对单个数据点的微小变化会产生截然不同的影响（清零不等于忽略，考虑因果与依赖）



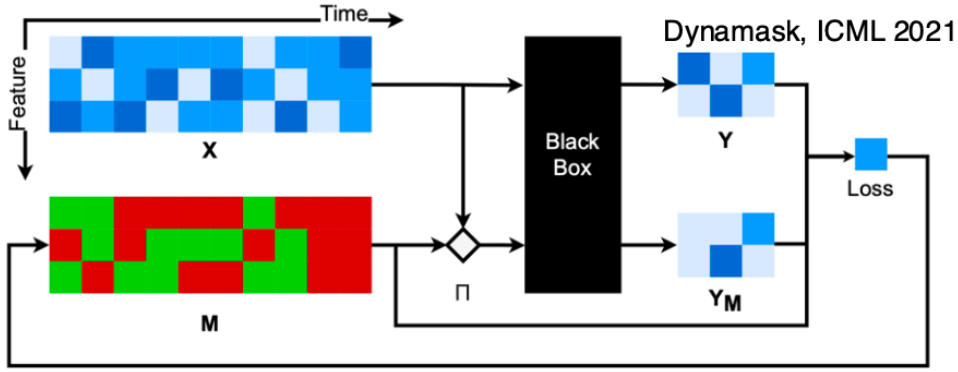
蓝色/灰色：大量样本的时序数据

红色：一个或几个原型序列

时序预测模型的解释可以通过展示最相似的几个原型，并指出这些原型在关键时间点的共同行为，来证明预测的合理性

可解释性方法的具体技术

- **现有时间序列解释器的不足**
  - **扰动是连续的：**
    - 时间序列具有内在的顺序和结构，对某点进行随机或孤立的扰动可能会破坏数据的自然形态和时间依赖性
  - **只提供基于实例的解释：**
    - 只能解释“为什么这个特定的序列会得到这个预测结果”，但无法将模式联系起来，缺乏全局观
  - **性能无法匹配通用解释器：**
    - 无法准确反映模型实际的决策逻辑
    - 对输入进行微小、合理的改变，解释结果却发生剧烈变化

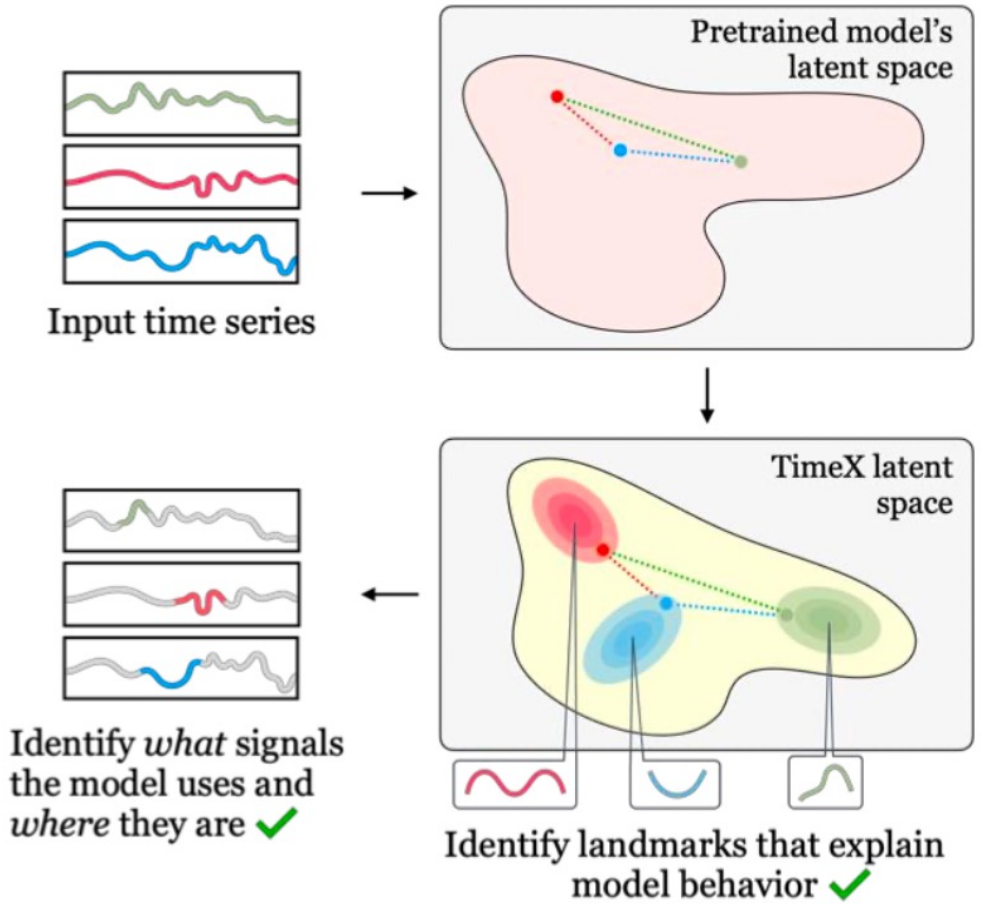


时间序列解释器的理想目标

- 解释结果必须尊重时间连续性，并且能够以人类容易理解的方式进行可视化
- 释器应该能够准确地定位时间序列中真正具有预测力的那段信号，并揭示其背后的潜在模式（周期性、波峰等）
- 能够从局部解释中提取全局的洞察，实现从单个案例到模型整体行为的过渡

可解释性方法的具体技术

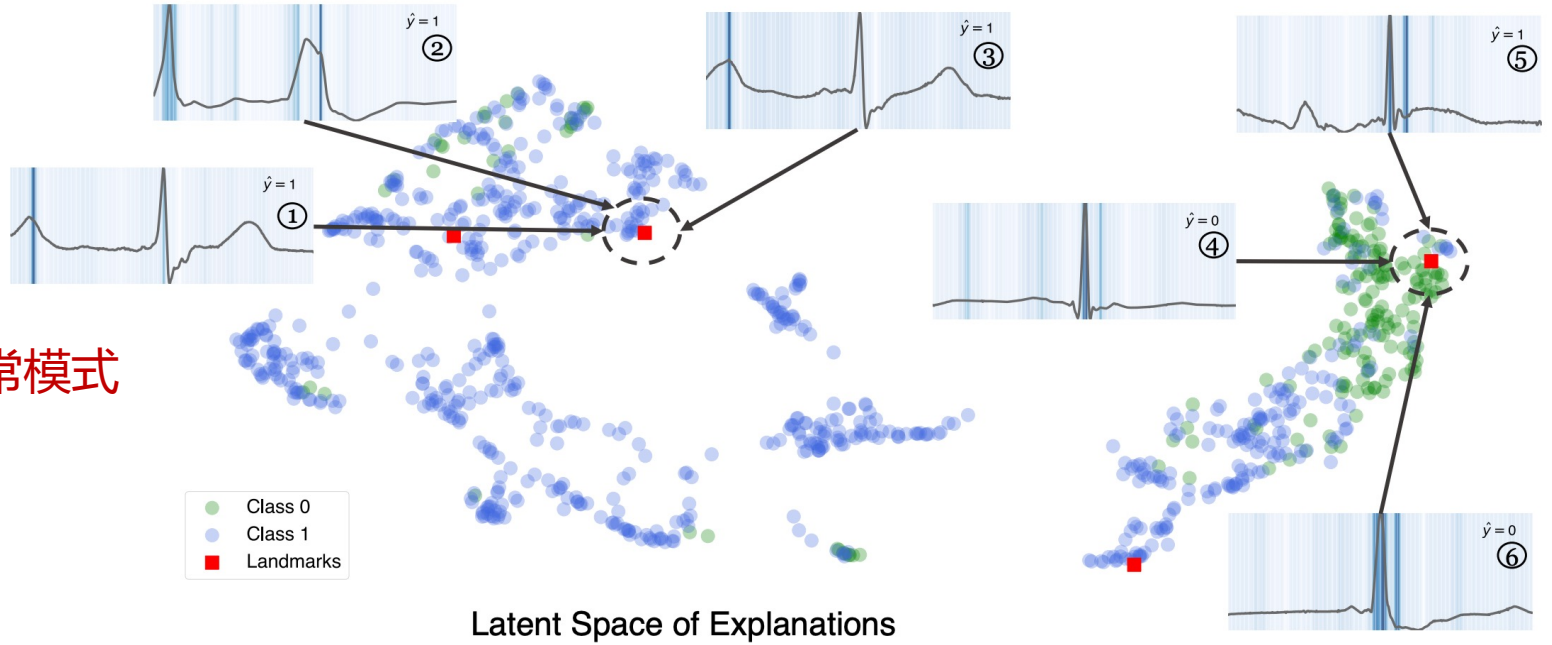
- **TimeX**：引入“一致性”来增强时序模型解释的质量
  - **代理模型**：使用一个代理模型来模仿一个预训练黑盒时间序列模型的行为
  - **对掩码样本进行推断**：通过遮盖或移除时间序列中的某些部分，来观察模型预测的变化
  - **模型行为一致性**：
    - **在潜在空间层面强制执行忠实性**：不仅要求解释模型在预测输出上与黑盒模型保持一致，还要求这种一致性在模型的潜在空间层面也得到保证，确保解释器真正捕捉到模型内部的深层逻辑



学习一个与黑盒模型一致的潜在空间，且可以定位解释行为，确保代理模型对黑盒模型的模仿是忠实且一致的

可解释性方法的具体技术

- **TimeX学习到的地标代表了时间序列中的重要模式：**
  - 地标将复杂的潜在解释空间划分成不同的区域，每个区域对应着一种可解释的特定模式
  - 通过观察信号与地标的关系，将抽象的潜在空间概念映射回具体的、可解释的时间模式
  - 直观理解模型基于哪种波形特征做出了最终的分类决策

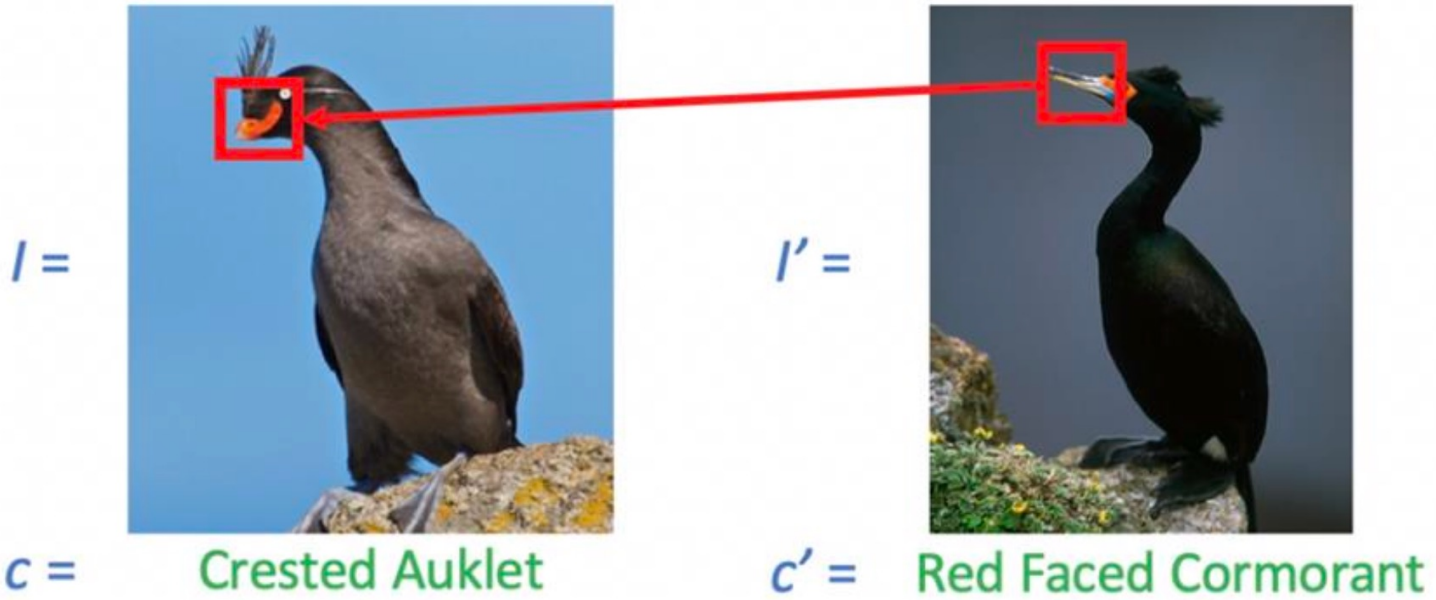


标签为1的正常模式

标签为0的异常模式

可解释性方法的具体技术

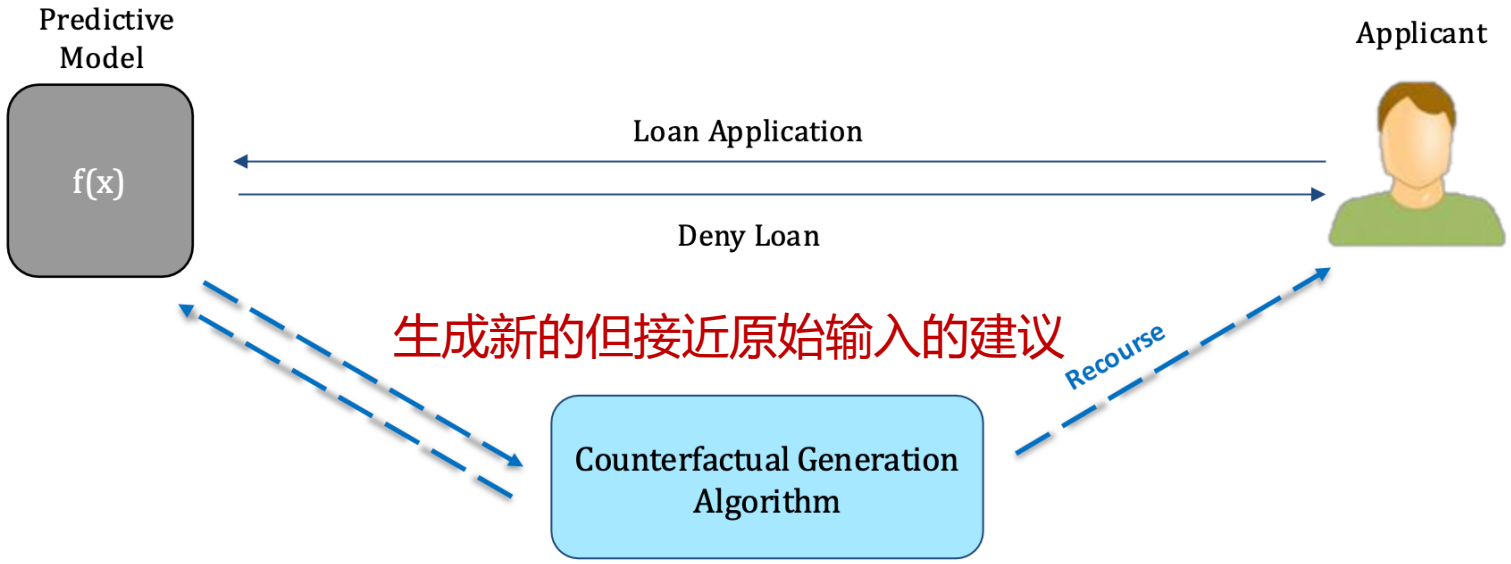
- **反事实解释方法：**
  - **思想：**为了让模型的预测结果发生改变，输入数据的哪些特征需要改变？需要改变多少？
  - **目标：**揭示模型认为对最终决策最关键的输入特征，如图像中的关键局部区域



可解释性方法的具体技术

- **反事实解释方法的一般框架：**
  - 方法：当模型做出一个不利的决策时，提供一个可操作的建议，告诉用户如何改变输入特征来获得不同的结果

意义：将黑箱模型的结果转化为用户可理解和可操作的指导，有助于缓解自动化决策带来的不公平感

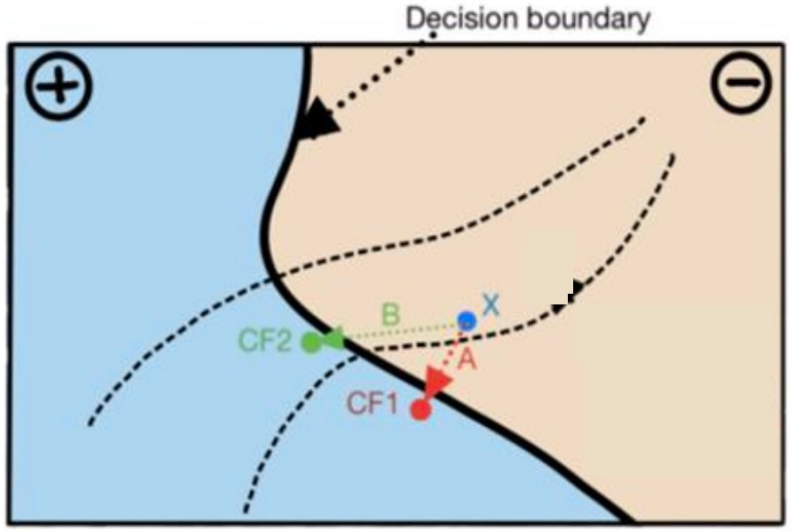


Recourse: Increase your salary by 50K & pay your credit card bills on time for next 3 months

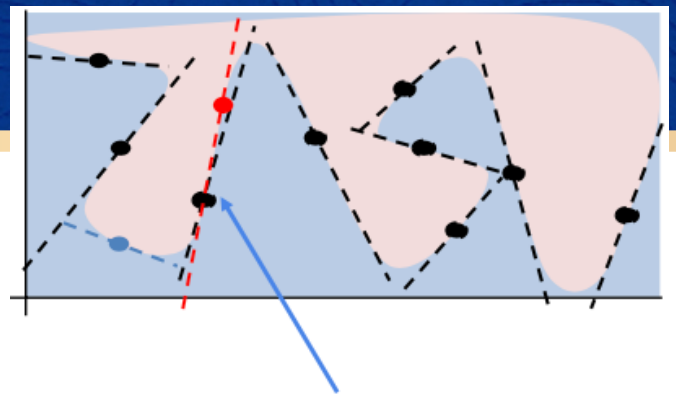
反事实解释：如果能实现这两个特定的、小的特征改变，当下次重新提交申请时，模型就会预测为通过

可解释性方法的具体技术

- **生成反事实解释：**
  - **直觉思路：**通过可视化一个决策边界（分类边界），并在边界附近找到原始数据点和反事实数据点，来展示反事实解释的本质，即找到使分类结果翻转的最小改变
    - 反事实数据点：满足可翻转结果、与原始点尽可能接近两个条件
  - **生成反事实解释的挑战：**
    - 如何评价反事实解释?
      - 最小化X到CF的距离、改变的特征数最少、CF符合常识
    - 生成反事实解释的算法如何理解预测模型?
      - 白盒方法：知道模型的结构和参数，允许使用梯度下降优化
      - 黑盒方法：只能通过输入和输出访问预测模型



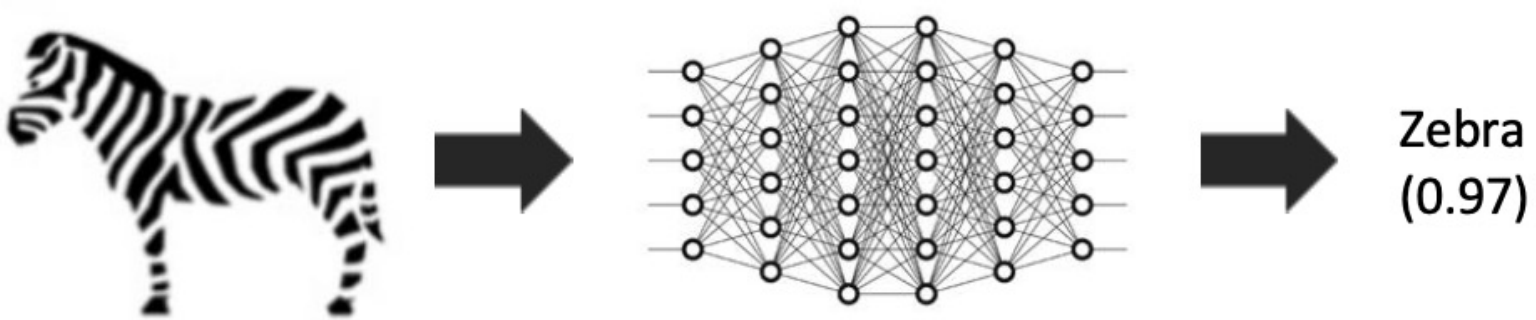
可解释性方法的具体技术



- **全局解释性模型——SP-LIME：**
  - **核心思想：**从大量的局部解释中，提取出最具代表性、最不冗余的一小部分，用以概括模型的整体行为
  - **LIME的局限性：**初衷是解释单个实例的局部预测行为，不可能展示针对每个数据点的所有局部解释
  - **SP-LIME 的设计目标：**让这 k 个被选中的局部解释能够有效地代表模型的全局行为
    - 局部解释的关键特性：
      - 代表性：选出的解释集合应该能够覆盖数据空间中不同区域的预测逻辑
      - 多样性：如果两个解释几乎一样，就只选择其中一个，以最大化信息密度
    - 核心机制：
      - 找到一个包含 k 个解释的集合，使得代表性得分和多样性得分最大化
      - 使用贪婪算法近似求解，即每次都选择当前能带来最大信息增益（即最大化代表性和多样性）的那个局部解释，直到选择了 k 个

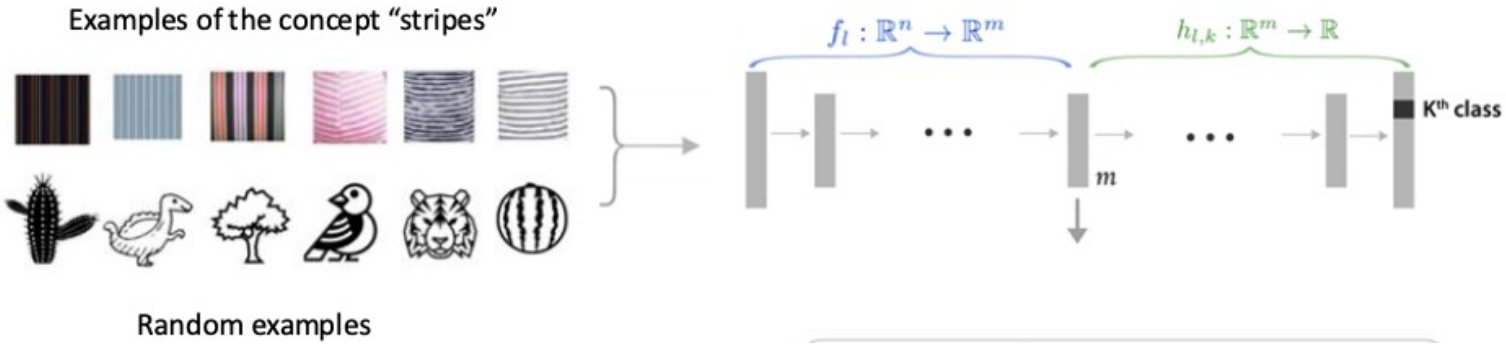
可解释性方法的具体技术

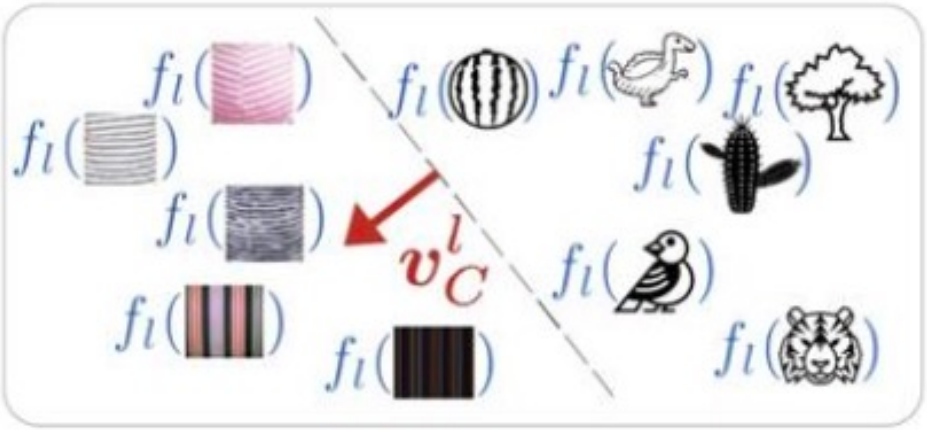
- **基于表示的解释方法：**
  - **核心思想：**从“像素”到“概念”的跨越
  - **问题：**传统的解释方法（如显著性图）通常只能展现图像中的哪些像素对预测最重要（比如高亮斑马身体部分的像素）。但像素是低层级特征，它不能体现模型是否真正理解了人类的高级概念（如“条纹”、“耳朵”或“四条腿”）
  - **方法：**不再关注输入层（像素），而是关注神经网络的内部表示层，从黑盒模型的内部向量中找到哪个对于某个高级概念（如条纹）的敏感度（导数）最高，这个向量叫做CAV（Concept Activation Vector）



可解释性方法的具体技术

- **基于表示的解释方法——TCAV：**
  - **由人类用户定位希望研究的概念**
    - 概念样本：如条纹
    - 随机样本：无关对照组
  - **提取内部激活值：**
    - 第l层的m维向量作为输入的高维表示
  - **训练分类器并获取CAV：**
    - 线性模型将概念与随机样本分开
    - 垂直于边界的向量为CAV
  - **计算概念的重要性**
    - 看实际样本预测分数相对于CAV的方向导数
    - 如果我们把这个图片的特征往“条纹”的方向推一推，模型认为它是“斑马”的概率会变大吗？





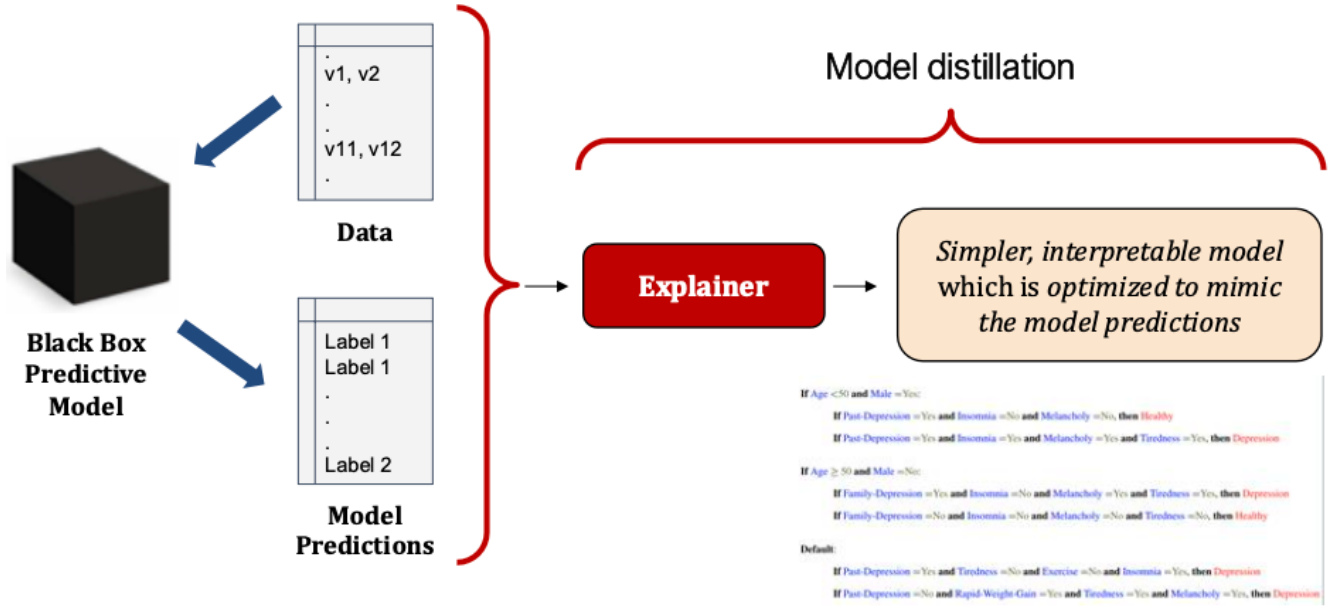
可解释性方法的具体技术

- **模型蒸馏解释方法：**
  - **思想：**用一个简单的、人类看得懂的模型，去模仿一个复杂的、人类看不懂的“黑盒”模型
  - **方法：**构建一个全局代理模型（global surrogate model）
  - **教师模型：**黑盒模型
  - **蒸馏过程：**
    - **关键点：不再关心数据的真实标签**

      只关心黑盒模型觉得它是什么

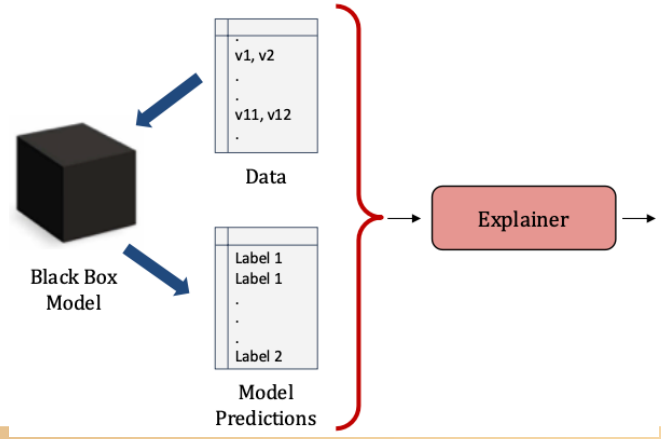
      即解释黑盒
    - 训练集：原始输入+黑盒模型预测结果
    - 输出：简单、可理解的学生模型

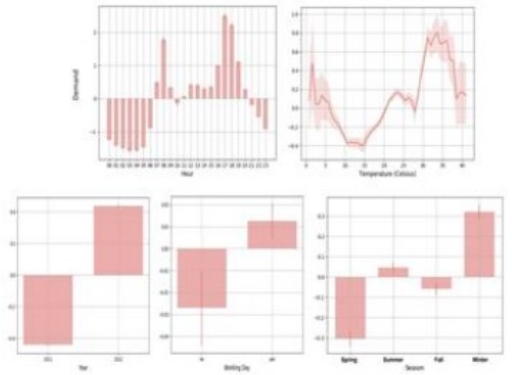
      如决策树、线性回归、规则集等
    - 优点：模型无关、结果直观；缺点：保真度问题



可解释性方法的具体技术

- **模型蒸馏解释方法：**
  - **常用学生模型：**决策树（Decision Trees）、通用加性模型（Generalized Additive Models）、决策规则集（Decision Sets）
  - **决策树：**将黑盒模型转化为规则路径，从最重要判断特征到最终预测结果，非线性切割的模拟能力强，符合人类的决策过程
  - **通用加性模型：**将黑盒模型的预测分解为每个特征独立贡献的叠加，$y = f_1(x_1) + f_2(x_2) + \cdots$，每个特征都有一个独立的函数来描述其对预测的影响程度，特征解耦分析，适合分析连续变量的趋势





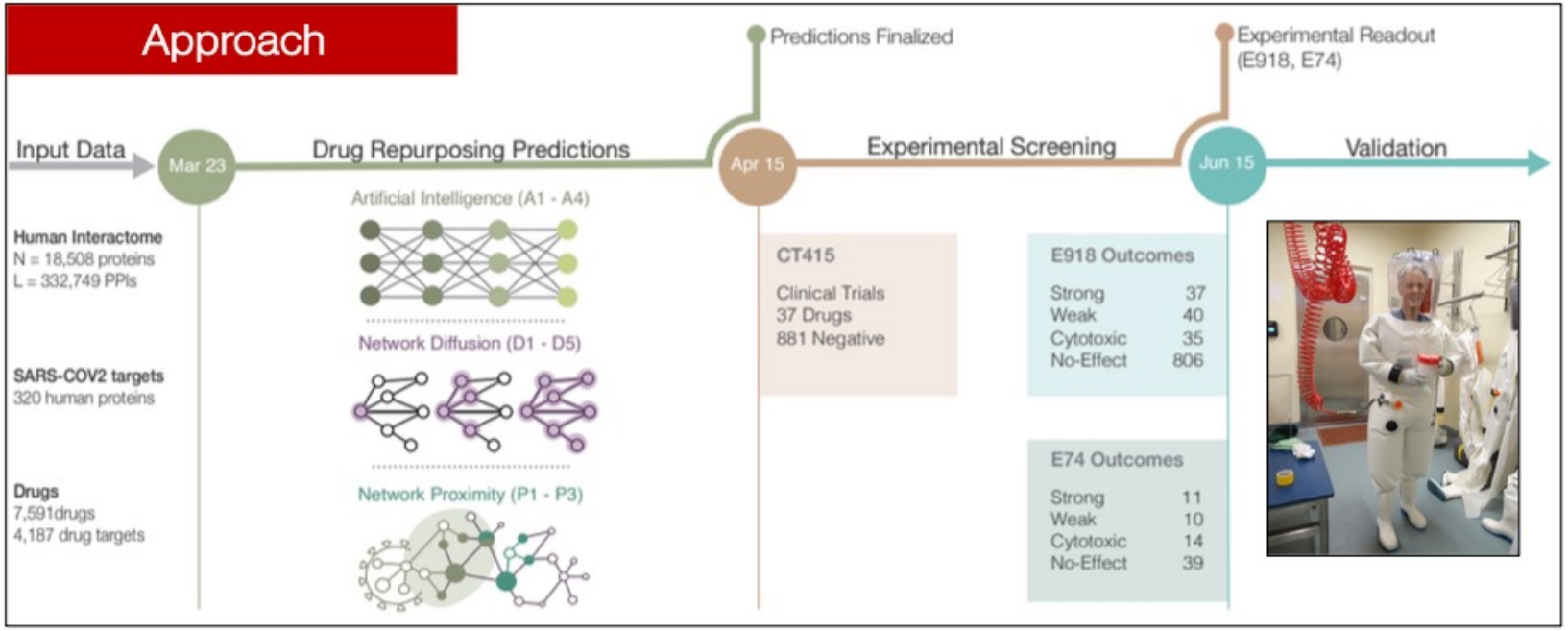
今日课程内容

- ✓ 什么是可信AI
- ✓ 对AI的预测结果进行解释
- 👉 医学场景案例分析
- AI公平性定义
- 公平AI的方法框架
- 个体公平与群体公平

案例：加速治疗创新

- **背景：**流行病要求以前所未有的速度开发出安全且有效的疗法
- **挑战：**传统的、迭代式的药物开发、实验和临床测试以及新药批准过程，通常需要数年时间
- **方法：**利用 AI 加速药物开发（特别是药物再利用/老药新用），有望将数年的时间压缩到数月甚至数周

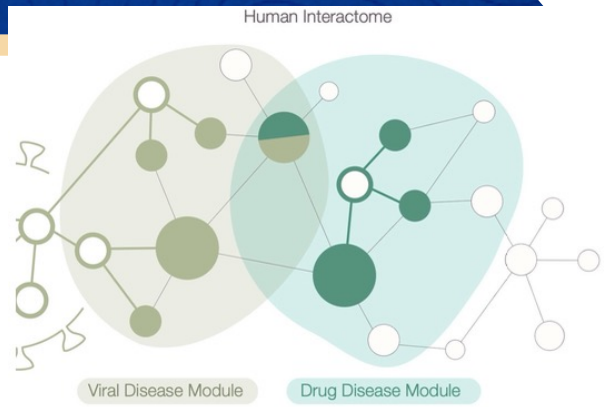
收集数据（蛋白质及相互作用）、疾病靶点、药物及其靶点

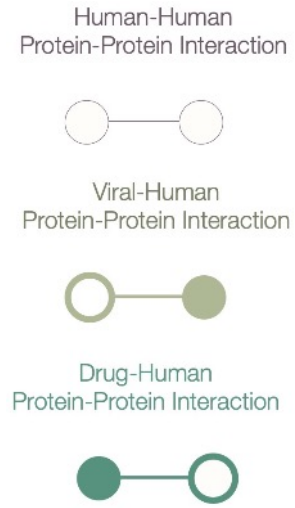


表现优异的药物进入最后体内与临床前验证阶段

AI模型+网络分析方法预测可对抗疾病的现有药物　通过实验筛选最优希望的药物

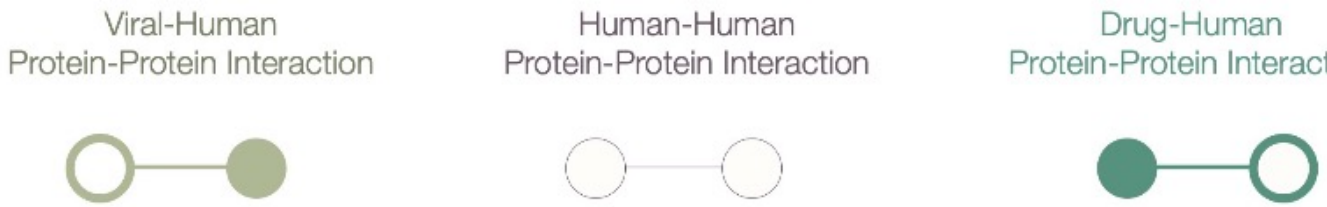
案例：加速治疗创新





- **通过网络分析设计治疗方案：**
  - **疾病的本质**：疾病的发生是由于基因的正常功能被破坏
  - **药物的作用**：药物的目的是通过干预来恢复这些被破坏的基因的功能，从而治疗疾病
  - **分析目标**：确定哪些**化合物**可以干预疾病
  - **分析方法**：生物网络建模与分析
    - 人类互作用组：PPI（人类蛋白互作用）、VPI（病毒蛋白与人类蛋白互作用）、DPI（药物或对应靶点与人类蛋白互作用）
    - 网络模块划分：病毒疾病模块（病毒直接或间接影响的基因或蛋白）、药物疾病模块（与潜在药物靶点相关联的基因或蛋白）
    - 分析目标：找到一个现有药物，通过其DPI能够有效**抵消或恢复**被**VPI** 破坏的功能
      - 寻找在网络中与病毒靶点**具有高接近度**或**高网络扩散影响**的化合物，有潜力修复被扰动的网络

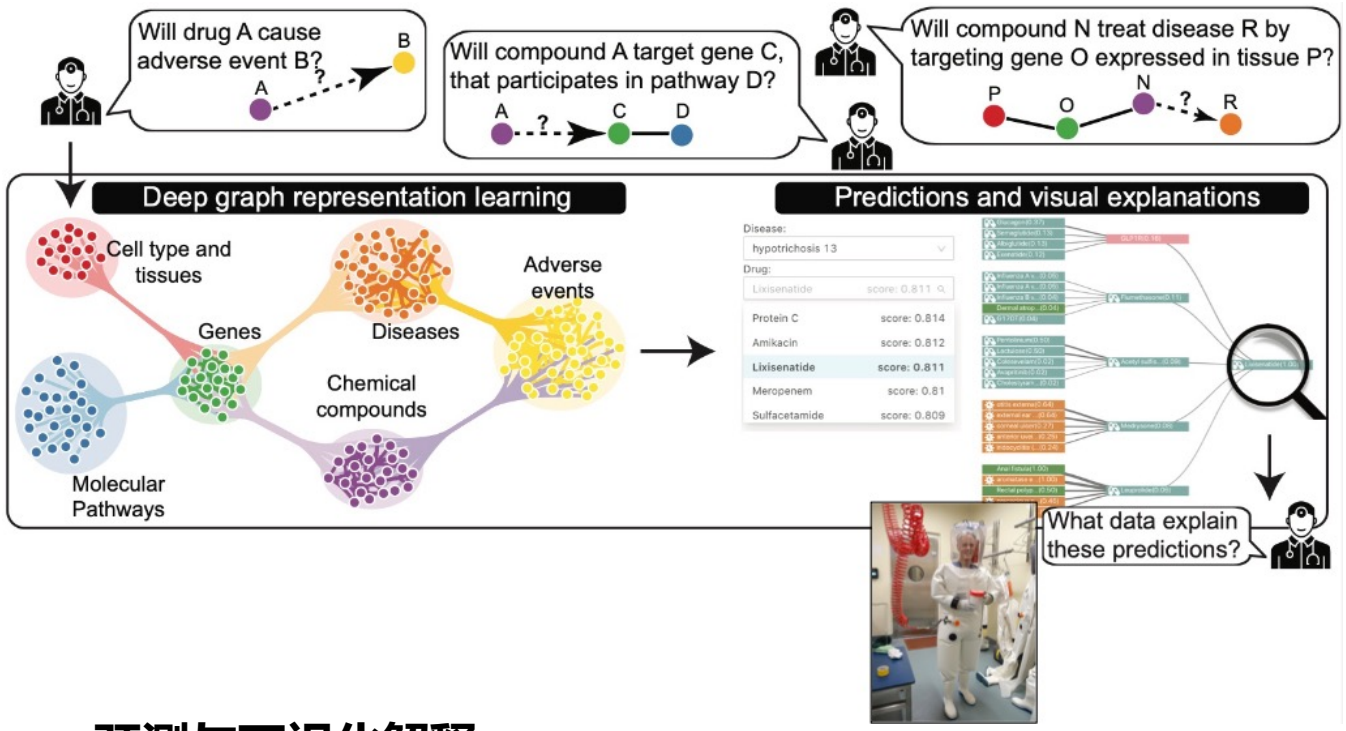
案例：加速治疗创新



- **数据集与实验设定：**
  - **构建 COVID-19 药物再利用知识图谱**
    - 表示人体内蛋白质之间的已知物理相互作用（PPI）
    - 已批准药物-人类蛋白质相互作用（DPI）
    - 疾病相关的蛋白质集合（VPI）
    - 已批准药物-疾病治疗方法（作为训练数据或已知阳性样本，让模型学习有效治疗的网络特征）
  - **机器学习任务建模**
    - 给定已批准药物-疾病治疗方法，识别 COVID-19 的候选治疗方案
      - 模型学习**已知有效药物-疾病**在知识图谱中表现出的**网络特征**
        - 如药物靶点与疾病相关蛋白质之间的**接近度**、**网络拓扑**或**扩散模式**
      - 模型预测未被批准用于 COVID-19 的药物中，哪些药物在网络特征上与**已知有效的治疗方案**最相似

案例：加速治疗创新

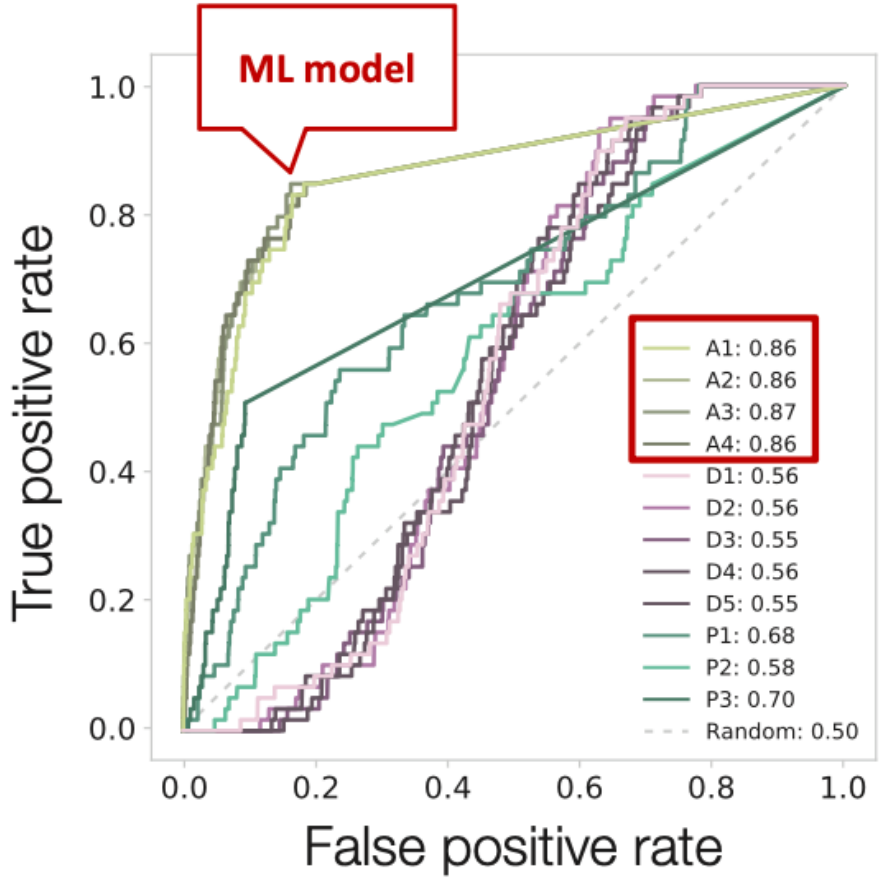
- **核心方法——图机器学习：**
  - 通过图推理回答一系列问题
    - 药物 A 是否会导致不良事件 B
    - 化合物 A 能否通过通路 D 中的基因 C 来治疗疾病
    - 化合物 N 能否通过在组织 P 中表达的基因 O 来治疗疾病 R
  - **知识图谱中实体：**分子通路、细胞类型与组织、基因、化合物、疾病、不良事件、…
  - **图表示学习：**学习每个实体节点的低维向量表示，使得相似的节点在向量空间中距离相近



- **预测与可视化解释：**
  - 使用学到的表示进行计算（如能够治疗相似疾病的药物在嵌入空间中会彼此接近）
  - 针对某个疾病给出多个药物的治疗可能性分数
  - 提供解释，知识图谱中哪些关系路径支持了这些预测

案例：加速治疗创新

- **针对COVID-19药物再利用的结果：**
  - **评估：**测试每种方法预测已注册用于 COVID-19 临床试验的药物的能力
  - **测试集：**67种已进入COVID-19临床试验的药物
  - **ROC曲线分析：**敏感性（True Positive）与特异性（False Positive），越靠左上角越好
  - **发现：**
    - 最佳性能由基于 GNN 的AI方法获得（0.86-0.87）
    - 第二佳性能由接近度方法提供（简单的网络分析指标）
    - 扩散方法表现较差（对于该任务，传播模式不够重要）



案例：加速治疗创新

- **试验筛选的结果：**
  - **待筛选化合物：**918个被模型预测出的化合物
  - **测试对象：**针对 SARS-CoV-2 病毒在人源细胞中的功效
  - **试验地点：**美国国家新兴传染病实验室
  - **筛选结果：**
    - 62%的化合物在人源细胞中显示出有效性
    - 传统的随机或未引导的筛选方法的命中率仅为0.8%
    - 比传统药物再利用方法相比，AI命中率也高出一个数量级

CRank	Drug Name
1	Ritonavir
2	Isoniazid
3	Troleandomycin
4	Cilostazol
5	Chloroquine
6	Rifabutin
7	Flutamide
8	Dexamethasone
9	Rifaximin
10	Azelastine
11	Crizotinib

17	Celecoxib
18	Betamethasone
19	Prednisolone
20	Mifepristone
21	Budesonide
22	Prednisone
23	Oxiconazole
24	Megestrol acetate
25	Idelalisib
26	Econazole

Predicted lists of drugs

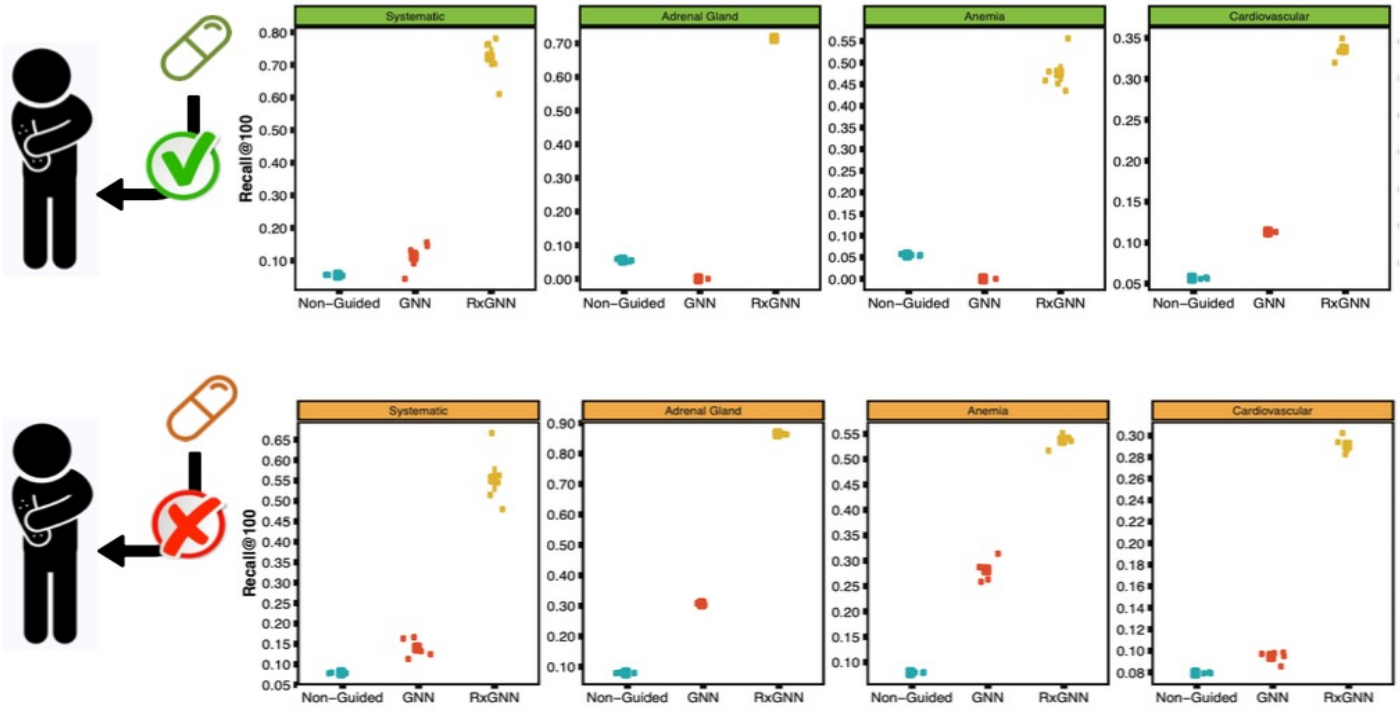
地塞米松：新冠重症有效药物

案例：加速治疗创新

- **预测治疗用途的方法对比：**
  - Non-Guided (非引导): 传统的基于随机或简单基准的方法
  - GNN (图神经网络): 使用图神经网络模型，但只使用了基本的图特征
  - RxGNN (增强型GNN): 经过特殊设计的、更复杂的图机器学习模型，例如专门用于药物再利用的框架

无论在有效治疗还是无效治疗方面，为药物再利用设计的图机器学习模型都能够有效地从知识图谱中学习到药物-疾病的复杂关联

预测哪些药物对于不同疾病治疗**有效**

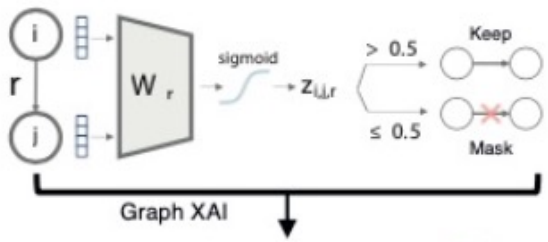


预测哪些药物对于不同疾病治疗**无效**

案例：加速治疗创新

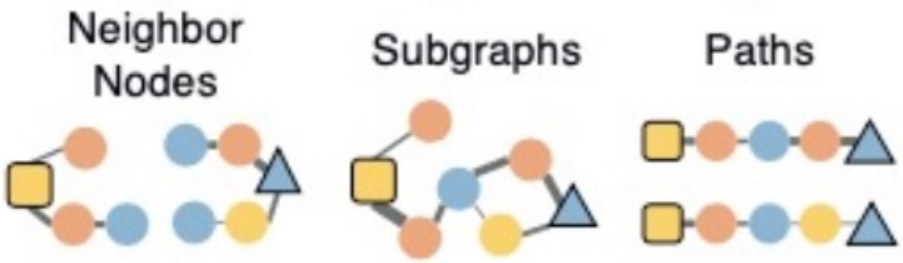
- **解释模型的预测：**
  - **核心思想：**告诉用户模型做出预测是基于知识图谱中哪些特定信息，找出一个最小和最具影响力的子图
  - **图可解释性（Graph XAI）通用框架：**
    - 基于**可微边掩码**（Differentiable Edge Masking）的 GNN 解释器

输入节点表示和可学习权重



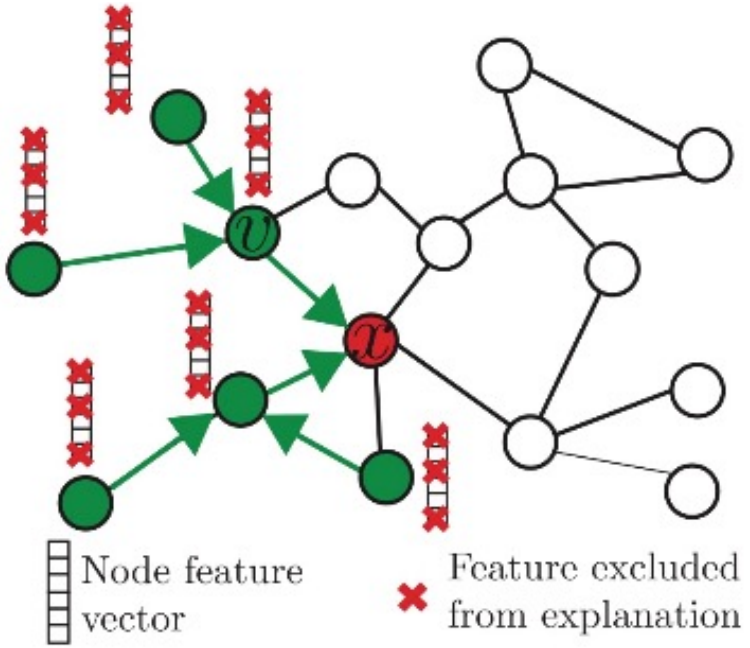
学习得到图上的掩码，使得被掩码保留下来的子图能够最大程度保留预测模型的预测结果，同时最小化被保留的边

- **解释的粒度：**
  - 最粗粒度：与目标节点直接相邻的邻居节点
  - 中等粒度：包含目标节点及其多跳邻居的子图
  - 最细粒度：从目标节点到其他关键节点的特定路径



案例：加速治疗创新

- **GNNExplainer：**
  - **核心思想：**为 GNN 模型对特定节点或链接的预测提供**事后 (post-hoc)** 解释，使预测过程透明化
  - **输入：**给定 GNN 对节点或链接 x 的预测 f(x)
  - **输出：**一个小的子图 M_x，以及一小部分节点特征作为解释，要求 M_x 必须是对预测**影响最大的**部分
  - **方法：**优化一个**掩码** M_x 来识别最重要的子图，即如果从图中移除一个节点 v，能**显著降低**原预测结果的概率，那么 v 就是一个很好的**反事实解释**
  - **效果：**提供一个**紧凑且高度相关的子图和特征集**作为解释，帮助用户理解 GNN 是如何做出特定决策的



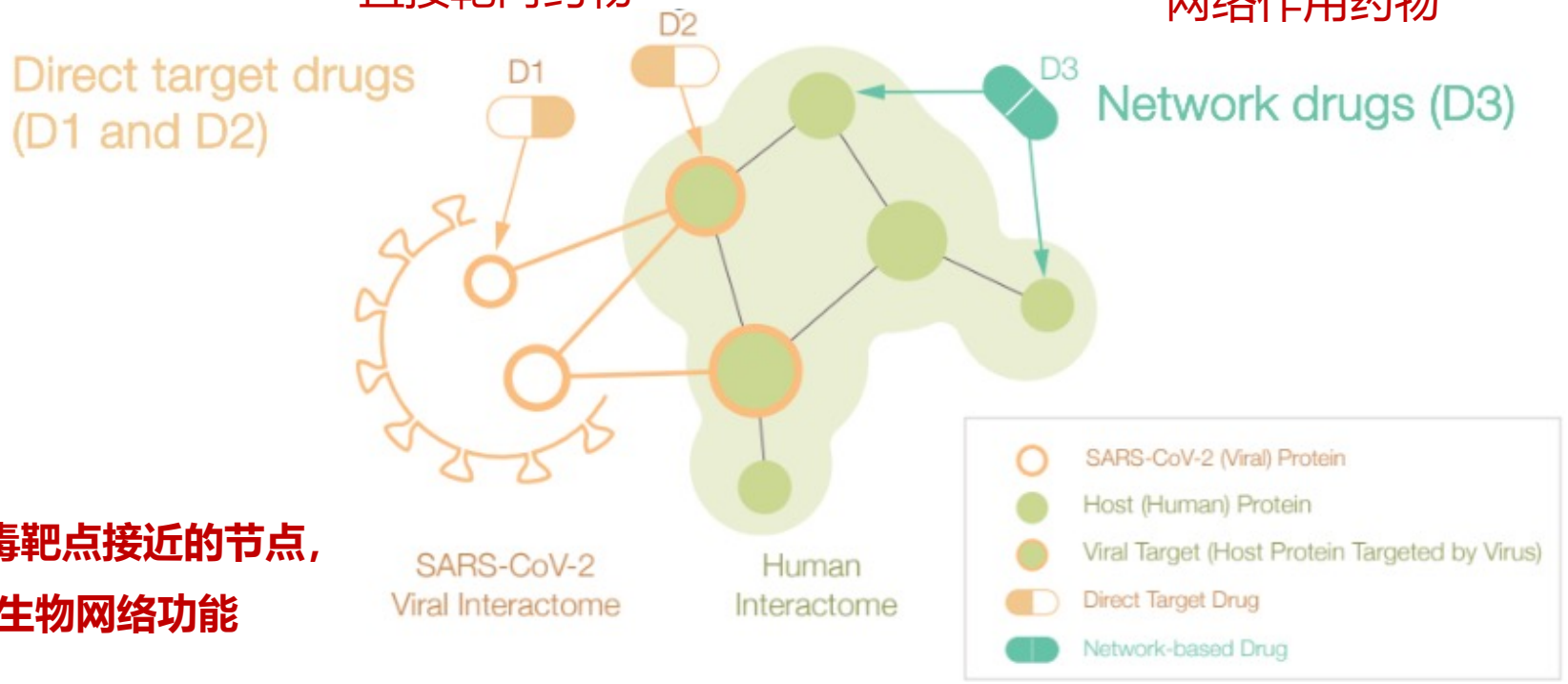
案例：加速治疗创新

- **对实际有效的药物解释其作用机制：**
  - **场景：**模型预测出有效药物后，科学家需要了解**为什么**这些药物有效，才能进一步推进临床试验
  - 关键发现：网络作用机制

- 77个预测出阳性的药物中有76个并不能直接与SARS-CoV-2 靶点结合
- 些药物依赖于基于网络的作用，**间接调节**人体自身的生物通路来达到治疗效果

直接靶向药物

网络作用药物



得出结论：一些药物可作用于PPI中与病毒靶点接近的节点，从而间接地或弥补性地恢复被病毒干扰的生物网络功能

浙江科技大学 ZHEJIANG UNIVERSITÄT FÜR WISSENSCHAFT UND TECHNIK

案例：加速治疗创新

- **DrugExplorer——基于 GNN 的药物再利用可视化解释：**

交互式控制：

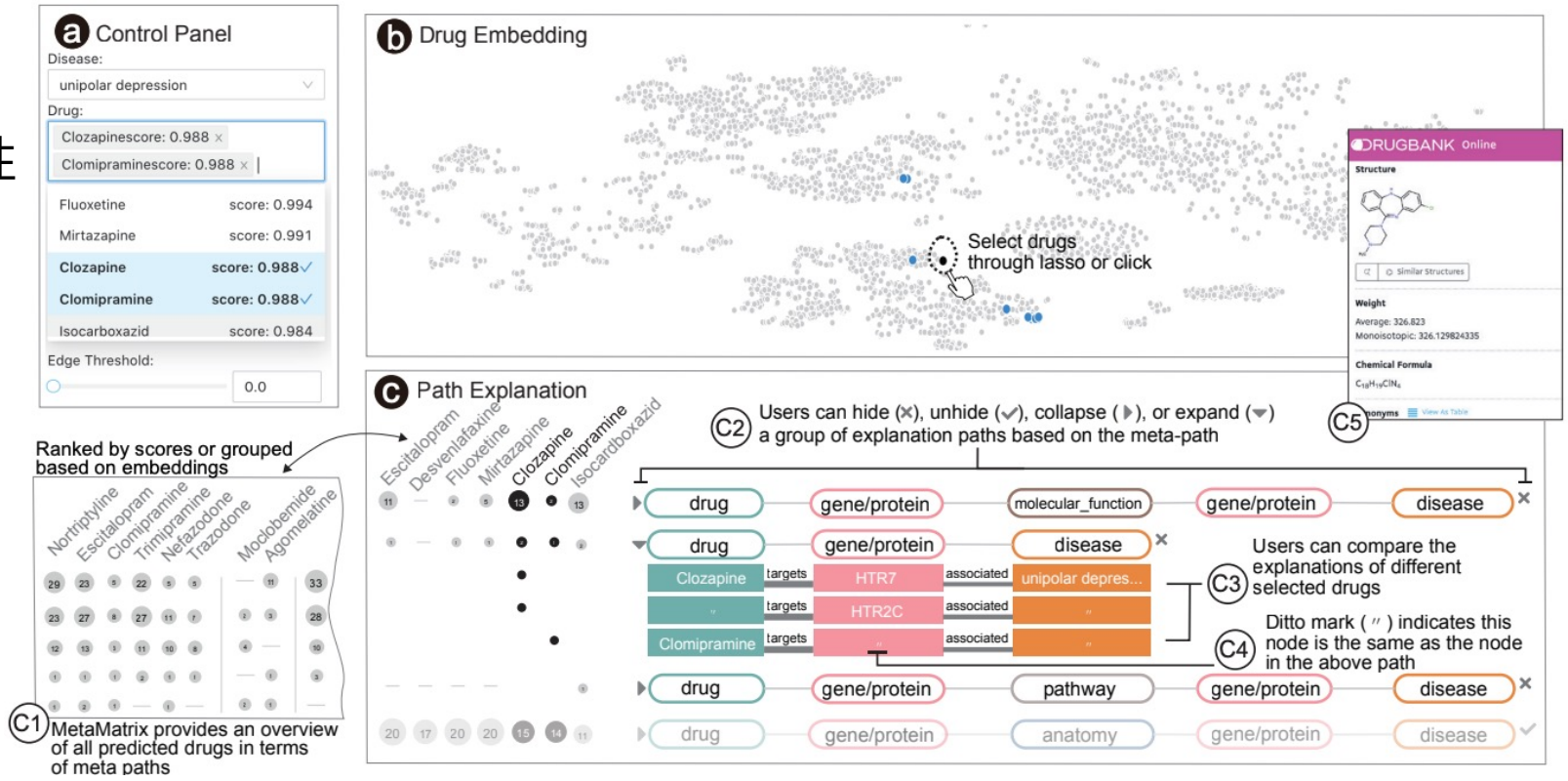
- 输入疾病，展示各种药物及其得分
- 边阈值：过滤解释路径中边的重要性

药物嵌入表示：

- 提供所有预测药物的概览和相似性
- 将药物的GNN嵌入投影到二维空间
- 可交互选择药物，展示药物信息

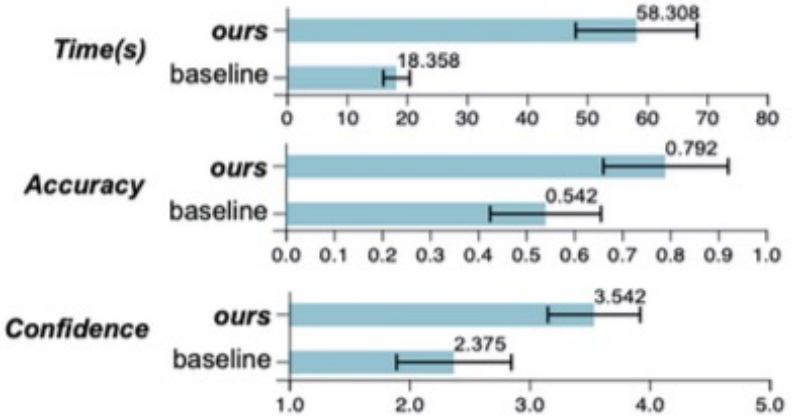
路径解释：

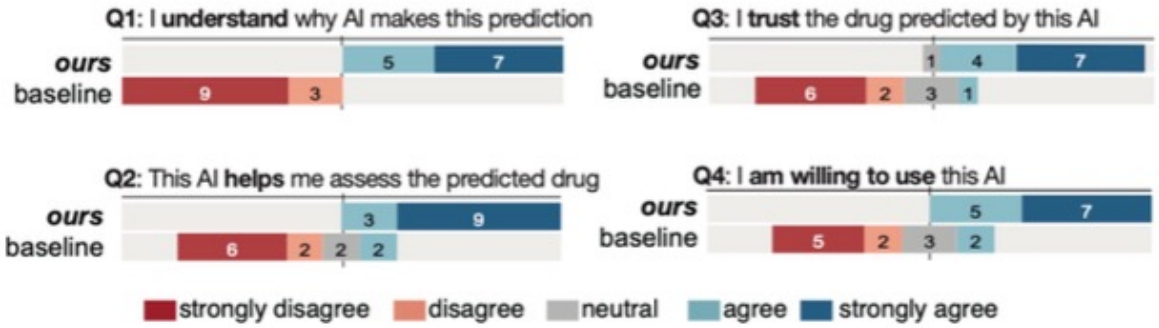
- 元路径概览，元路径重要性
- 展示用户选择药物的具体路径
- 对比不同药物解释路径的相似程度



案例：加速治疗创新

- **DrugExplorer——临床医生为中心的研究：**
- **定量指标对比：**
  - 与无解释的基准方法对比，临床医生探索和得出结论的时间减少68%
  - 临床医生使用解释工具时对药物预测问题的回答准确率提升46%
  - 临床医生系统回答的置信度更高
- 用户满意度定性评估：
  - Q1: 我理解为什么 AI 能做出这个预测（提升理解）
  - Q2: AI 可以帮助我评估预测的药物（批判性评估）
  - Q3: 我信任这个 AI 预测的药物（提升信任度）
  - Q4: 我愿意使用这个 AI （实用性和临床接受度）





AI解释性中存在的实践和伦理挑战

- **Q1: 决策者是否从解释中受益?**
  - 现实世界效益的证据是复杂的，解释是否能带来预期的决策质量提升，取决于任务、用户的专业水平和解释的质量
- **Q2: 得到的解释如何与对 AI 的信任对齐?**
  - 人们往往更倾向于相信带有解释的 AI，即使这些解释是误导性的或不完整的，如果解释是片面的或经过精心设计的，它们可能会导致用户**过度信任**存在缺陷的 AI 系统
- **Q3: 得到的解释如何与对公平性的感知对齐?**
  - 解释可能被用来**合理化**模型的不公平决策，如果一个模型做出有偏见的决定，但提供了一个看似合理的解释，用户可能会认为该决定是公平的，尽管它在统计上或伦理上存在问题
- **Q4: 对抗者能否欺骗解释算法，进而欺骗用户?**
  - XAI 算法本身是机器学习模型，它们也容易受到**对抗性攻击**，攻击者可以微调输入或模型本身，使得模型做出不想要的预测，但同时生成一个**看似无辜或合理的解释**

今日课程内容

- ✓ ■什么是可信AI
- ✓ ■对AI的预测结果进行解释
- ✓ ■医学场景案例分析
- 👉 ■AI公平性定义
- ■公平AI的方法框架
- ■个体公平与群体公平

在高风险领域采用AI

- **高风险领域：** AI 的决策直接关系到人的生命、自由、健康和经济利益
  - 医疗与生命科学：误诊或错误建议导致患者死亡
  - 刑事司法：算法存在偏见导致不公正裁决
  - 经济活动：AI的群体歧视导致经济机会被剥夺
  - 社会福利：错误的算法导致社会保障的不公平
- **揭示现象：** 要获得相同的医疗救助（模型预测高风险），黑人必须比白人病得更严重（实际患者数）
- **原因：** 训练数据中，算法对花钱少和比较健康产生了关联，而同样健康状况下，黑人的医疗花费通常低于白人
- **意义：** 需要打开黑盒模型，审视决策逻辑，确保没有偏见

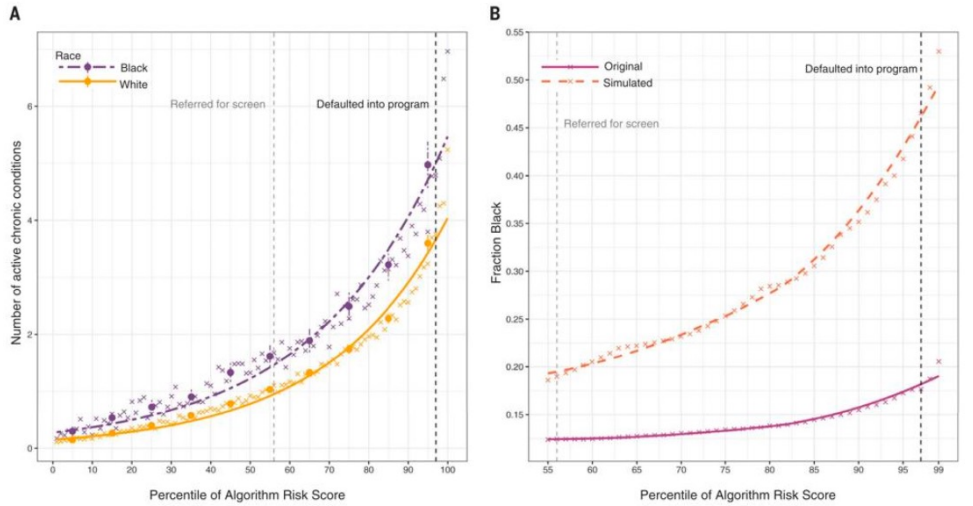


Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated"; at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

Obermeyer et al. *Science* 2019

导致AI不公平的案例

- **高风险医疗管理：在大型医疗系统使用商业预测模型来决定哪些病人需要额外的护理资源**
  - 发现在相同的风险分数下低收入患者实际上比高收入患者病得重得多
  - 原因：变量特征错误关联（医疗费用特征权重高）——目标错误
- **刑事风险评估：法院使用 AI 给被告打分，预测他们再次犯罪的风险**
  - 分数直接决定了被告能否被保释、判刑多重以及能否获得假释
  - 原因：模型基于历史定罪数据训练，数据收集存在偏向性，模型放大了社会偏见——数据偏颇
- **人脸识别系统：用于监控摄像头、智能手机解锁，甚至自动驾驶汽车**
  - 系统在特定人群身上的表现很差，错误率极高
  - 原因：训练数据不平衡导致对部分群体出现脸盲，用于自动驾驶存在安全隐患——数据缺失

COMPAS的算法偏见

- **COMPAS：用户评估罪犯在被释放后再次犯罪的可能性系统**
  - 法官会参考其给出的分数来决定是否批准保释、判刑多重以及是否给予假释
- **真实问题：大量使用案例发现其对黑人存在系统性偏见**
  - 黑人被告被算法标为“高风险”但他实际上没有再犯，白人被告被打为“低风险”但他实际上再次犯罪
- **在使用AI时的受保护群体特征：**
  - 种族、性别、年龄、宗教、国籍/身份、怀孕状态、残疾状态、基因信息 —— 不可变特征
- **使用AI受监管的领域（美国）：**
  - 信贷（平等信贷机会法案）、教育（民权法案）、就业（民权法案）、住房（公平住房法案）
- **对AI开发者的启发：**
  - 输入特征是否包含这些列？输入特征是否间接与敏感特征关联？模型是否涉及高风险领域？

机器学习模型的公平性

- **关于公平性的核心认知：**
  - 并非出于恶意：偏见发生的原因不一定是工程师或医疗人员的恶意，即使所有人都抱着最好的意图，偏见仍可能发生（不小心引入的间接关联变量）
  - 并非一劳永逸：即使一个算法现在没有偏见，也不代表它未来没有潜在偏见，公平性需要持续监测
  - 并非新事物：研究人员在过去数十年中一直关注和提出了对偏见的担忧和应对策略
- 公平性定义的角度：
  - 区别对待：模型直接使用受保护属性（如种族、性别）作为决策依据，从而对不同群体进行不同的处理
  - 不同影响：模型即使没有直接使用受保护属性，但对不同群体造成了不平等的、有差别的负面影响
- 偏见的系统性根源：
  - 数据采集过程中的偏见：如群体数据的缺失
  - 算法设计中的偏见：如选择了错误的优化目标
  - 模型部署/实施中的偏见：如认为设定了不合适的阈值参数

机器学习模型的公平性

- **如何评价机器学习模型是公平的：**
  - **基于群体的统计学定义：**模型在受保护群体之间的统计指标保持一致
    - 群体公平性：在所有群体中，得到积极结果的比例是相等的（人口均等性）
    - 校准性公平：对于任何一个给定的预测风险分数，所有群体中的实际结果发生率都必须相同
    - 错误率平衡：要求模型在不同群体中的错误率（False Positive / False Negative）保持一致
  - **基于数据和设计的定义：**关注模型输入数据和设计过程本身是否公平
    - 无意识公平：将受保护属性从训练数据中剔除（简单但容易产生间接关联）
    - 代表性公平：训练数据集本身对所有受保护群体具有充分的代表性
  - **基于个体的理想定义：**确保模型对相似个体给出相似的对待，而与群体身份无关
    - 个体公平性：两个在所有相关特征上相似的个体，无论其受保护属性如何，都应得到相似的预测结果
    - 反事实公平：如果将一个人的受保护属性（如种族）进行反事实假设的改变，而其他特征不变，模型的预测结果应该保持不变

机器学习模型的公平性

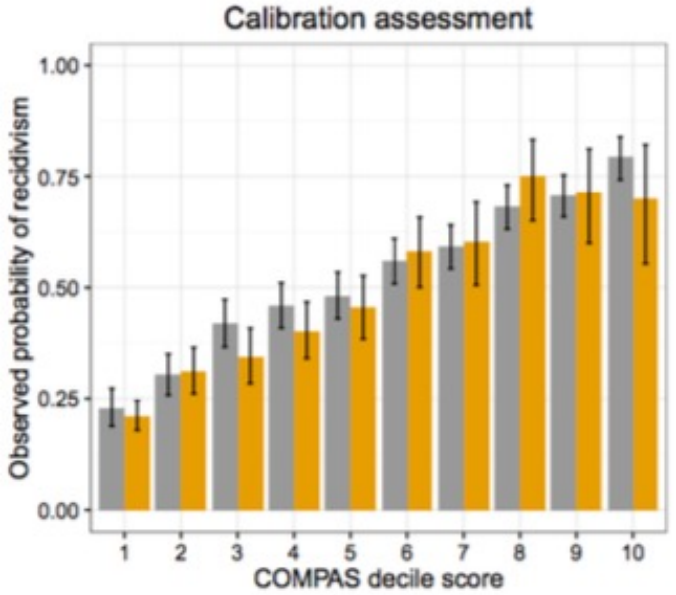
- **无意识公平（Fairness through Unawareness）：**
  - **原则：**不纳入受保护属性（种族、性别、宗教等），也不在算法中使用它们
  - **形式：**基于非敏感特征X来预测$\hat{Y}$，即使用条件概率$P(\hat{Y}=Y|X)$，而非纳入敏感群体S，使用$P(\hat{Y}=Y|X,S)$
  - **优点：**保证模型不会直接根据受保护属性做出判断
  - **缺点：**无法避免代理变量（Proxy Variables）
    - 数据中包含的其他特征（如地址或病史）仍可能与受保护属性高度相关
    - 算法可通过代理变量来“推断”或学习敏感属性，从而间接导致歧视（地址与社会地位相关）

机器学习模型的公平性

- **群体公平（Group Fairness）：**
  - **原则：**要求模型对所有受保护群体给出积极预测结果的比例必须相同
  - **形式：**预测结果 $\hat{Y}$ 为积极标签（如“健康”）的概率，在敏感群体 $S=1$ 和 $S=0$ 中必须相等，即 $P(\hat{Y}=1|S=1)=P(\hat{Y}=1|S=0)$
  - **优点：**直观上是平等的，从结果上保证了每个群体得到了平等的对待
  - **缺点：**与现实世界相冲突
    - 群体在现实中的基础概率本身存在差异，完美的分类模型也无法满足公平性要求
    - 不控制错误率：只要求最终的预测比例相等，但它不控制错误率，即不同群体中虽然积极标签比例相同，但是其中一群体预测错误过多

机器学习模型的公平性

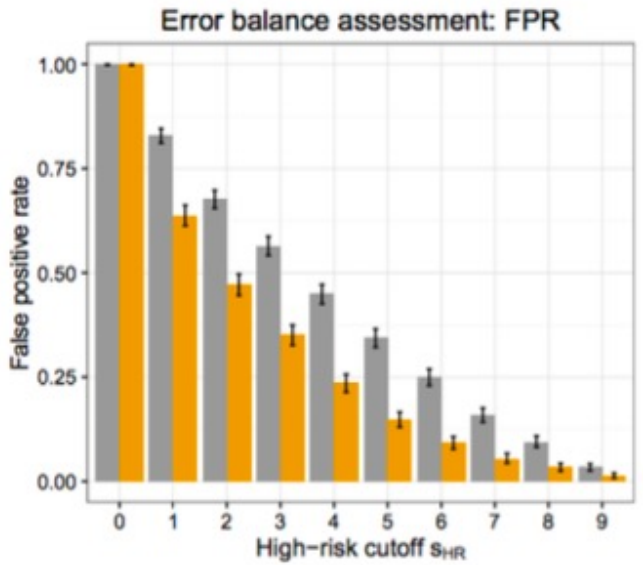
- **校准性公平（Calibration）：**
  - **原则：**要求在所有群体中，相同的预测风险分数对应相同的实际结果发生率（关注实际发生）
  - **形式：**真实结果Y为积极标签（如“健康”）的概率，在相同的预测分数r和不同群体$A = 1$和 $A = 0$ 中是相等的，即$P(Y = 1|R = r, A = 1) = P(Y = 1|R = r, A = 0)$
  - **优点：**确保在相同分数下，模型在不同群体中是代表相同的真实结果的
  - **缺点：**与错误率平衡不兼容
    - 不同群体基础概率不同
    - 满足校准性就必然导致错误率在不同群体中不一致



在任何给定的风险分数下，两个群体的实际再犯罪概率大体相同（即校准了）

机器学习模型的公平性

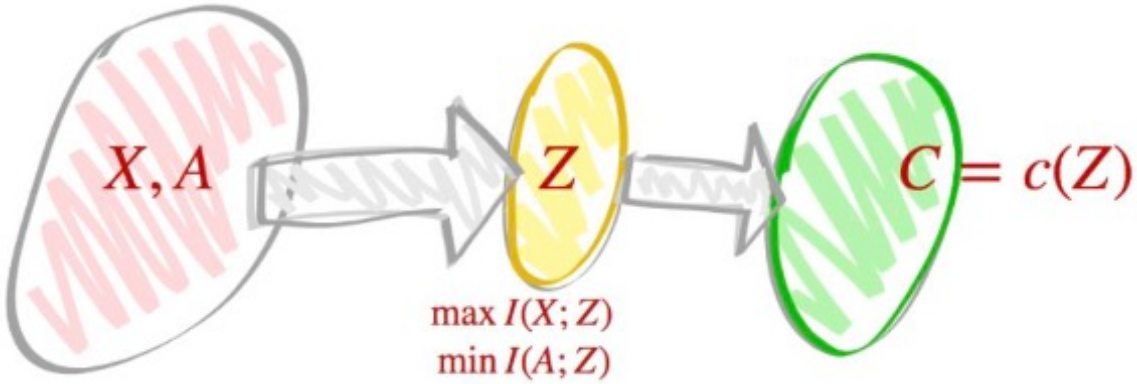
- **错误率平衡（Error rate balance）：**
  - **原则：**要求在所有受保护群体中假阳性概率（预测为积极，实际为消极）必须相同
  - **形式：**预测为积极结果$\hat{Y}=1$的概率，在真实结果为消极$Y=0$且属于不同群体（$S=1$和$S=0$）的情况下必须相等，即$P(\hat{Y}=1|Y=0,S=1)=P(\hat{Y}=1|Y=0,S=0)$
  - **优点：**确保了模型在不同群体中，将“好人”错误地判为“坏人”的概率是相同的（同样犯错）
  - **缺点：**
    - 与校准性不兼容，错误率平衡几乎总是与校准性不兼容
    - 不控制False Negative（预测是消极，但真实结果为积极）
      - 导致双重不平衡



在任何给定的风险分数下，两个群体的错误率不相同（即不满足错误率平衡）

机器学习模型的公平性

- **代表性公平（Representational Fairness）：**
  - **原则：**通过将原始输入特征向量 X 转换为“公平的表示” Z 来实现公平性
  - **形式：**模型尽量最小化 Z 中所包含的群体敏感性信息，同时尽量最大化 Z 中对最终决策有用的信息
    - 即找到一个中间表示 Z，使得 Z 与原始数据 X 保持最大的互信息 I(X; Z)，但与敏感属性 A 保持最小的互信息 I(A; Z)
  - **优点：**在保留重要预测信息的同时，减少提供给模型关于群体的敏感信息，可解决代理变量问题
  - **缺点：**
    - 必须在准确性和公平性之间进行权衡取舍
    - 过度消除群体信息导致模型损失必要的预测能力



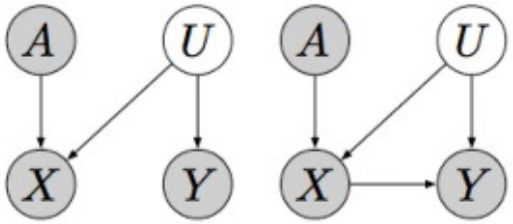
一种预处理或模型内处理方法

机器学习模型的公平性

- **反事实公平（Counterfactual Fairness）：**
  - **原则：**更改一个个体的敏感信息，而其他非敏感信息保持不变，模型预测结果应该保持不变
  - **形式：**基于所有未观察到的混杂因素 U，将 A 的值从 a 改变到 a′ 之后，预测值 $\hat{Y}$ 保持不变，即

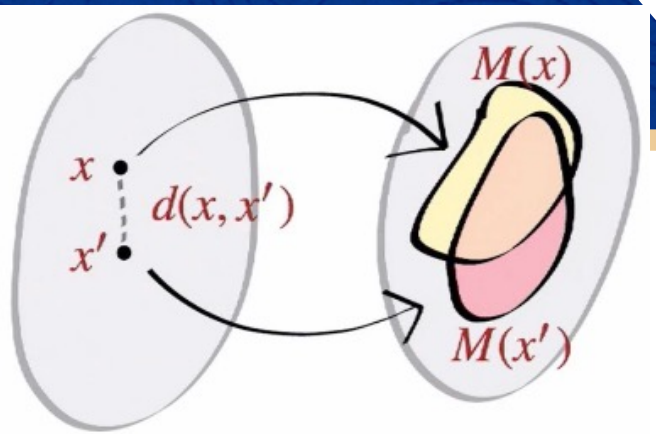
$$P\left(\hat{Y}_{A\leftarrow a'}(U)=y \middle| X=x, A=a\right)=P\left(\hat{Y}_{A\leftarrow a}(U)=y \middle| X=x, A=a\right)$$

  - **优点：**
    - 解决代理变量问题，不仅要求模型不看 A，还要求模型不通过 A 导致的任何潜在预测
    - 明确依赖关系，可以通过因果图来明确建模特征之间的依赖关系
  - **缺点：**
    - 因果图的问题，无法完美代表复杂现实，未观察到因素的影响



$$P(\hat{Y}_{A\leftarrow a}(U)=y \mid X=x, A=a)$$
$$=P(\hat{Y}_{A\leftarrow a'}(U)=y \mid X=x, A=a)$$

机器学习模型的公平性



- **个体公平（Individual Fairness）：**
  - **原则：**相似的个体应该被相似地对待，即如果两个人 x 和 x' 在所有相关特征上都相似，那么他们的模型预测结果 M(x) 和 M(x') 也应该非常相似
  - **形式：**如果两个个体之间的距离 d(x, x′) 很小，那么模型输出 M(x) 和 M(x') 也应该非常接近
  - **优点：**
    - 个体异质性建模，避免了将整个群体视为同质实体所带来的弊端，确保模型在对个体进行决策时，使用的是他们的个人特征而非群体身份
  - **缺点：**
    - 相似性定义困难：在高维数据中确定哪些特征是相关的或不相关需要深刻的领域知识

机器学习模型的公平性

- **公平性缺乏一个统一的定义和框架**
  - **核心问题：**
    - 研究内容分散：不同的定义（如群体公平性、校准性、错误率平衡、个体公平性等）之间存在差异和联系，使得研究人员和实践者难以把握关键点
    - 缺乏通用语言或框架： 不同的研究团体在公平性定义上未能达成共识
  - **学术观点：**
    - “没有人找到一个被广泛接受的良好公平性定义，就像我们对随机数生成器的安全性有共识那样”
    - “研究一个真正的公平性定义并非一个有成效的方向，因为技术上的考量无法裁决道德辩论”
    - “在公平性定义的问题上，许多定义和研究团体并未达成一致”
  - **总结：公平性不仅仅是一个技术问题，它是一个道德和哲学问题**
    - 技术无法提供一个万能的解决方案（如 无法同时满足校准性和错误率平衡）
    - 在高风险领域选择哪种公平性定义是一个社会和政策的决定，而非算法可以自行解决

今日课程内容

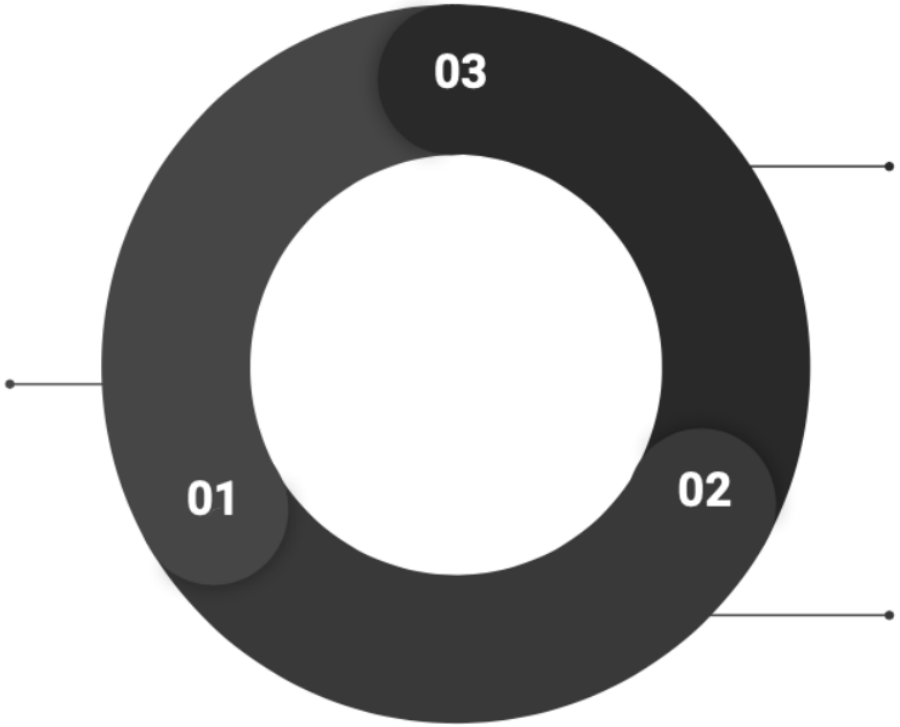
- ✓ ■什么是可信AI
- ✓ ■对AI的预测结果进行解释
- ✓ ■医学场景案例分析
- ✓ ■AI公平性定义
- 👉 ■公平AI的方法框架
- ■个体公平与群体公平

公平AI的框架

数据监管者

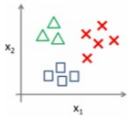
- 确定公平性标准
- 确定数据来源
- 审计结果

AUTHORITY



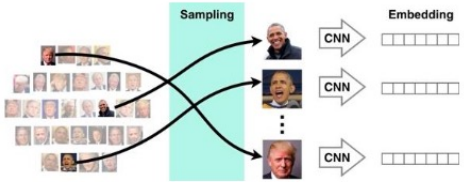
数据用户

- 构建或训练 ML 模型
- 预测或决策



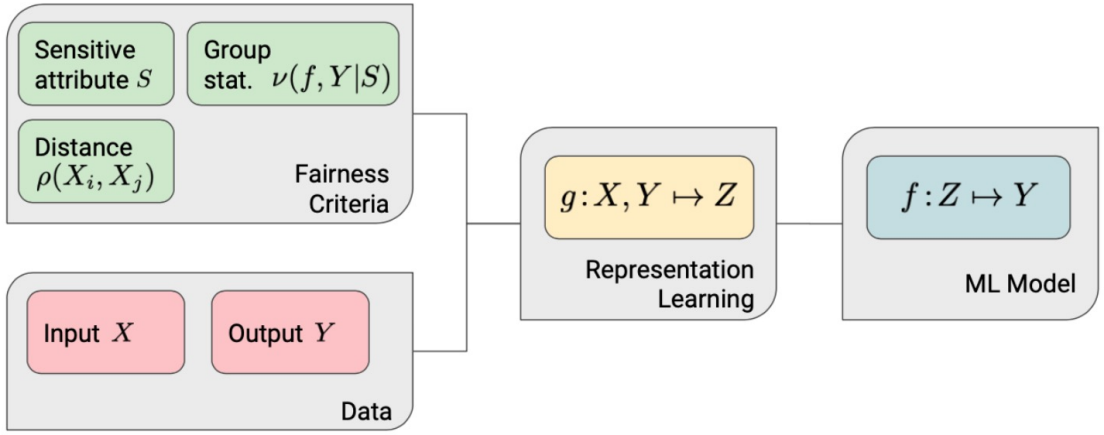
数据生产者

- 计算数据的公平表示
- 实施数据脱敏

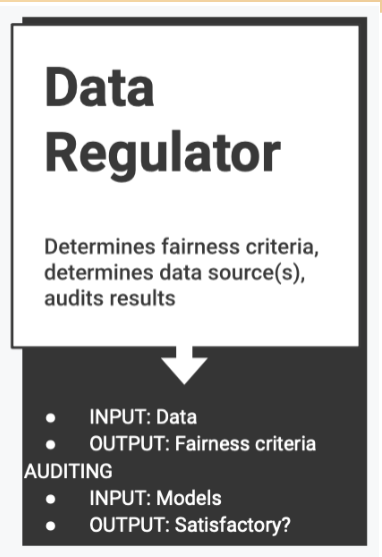


公平AI的框架

- **数据监管者 (Data Regulator)：确定公平性度量标准，并审计结果**
  - 确定敏感属性S：如种族、性别、年龄，不应影响模型的决策公平性
  - 计算群体差异统计量：衡量不同敏感属性群体S间的模型输出Y/f的统计差异
  - 差异距离定义：定义个体之间的差异距离
- **数据生产者（Data Producer）：负责执行表示学习以创建公平的特征**
  - 学习一个映射函数，将原始输入和输出转换为新的特征表示，该表示必须满足数据监管者定义的公平性标准，对敏感属性脱敏，防止下游模型利用属性进行有偏见的预测
- **数据用户(Data User)：使用公平特征训练最终的模型**
  - 训练一个预测模型，仅使用公平表示来预测输出，
  - 不担心原始数据中的敏感属性，只关注模型的预测性能



数据管理者



- **核心职责: 数据监管者是定义和监督公平性标准的权威方**
  - 定义公平性: 数据监管者确定要使用的公平性标准
  - 审计: 审计模型结果，以确保公平性标准得到满足
- **确定公平性标准:**
  - 输入：与用户、专家、法规、政策的互动，非纯技术问题，需要反映社会、法律和伦理的期望
  - 输出：公平性标准：规则集，例如： “模型在性别 S 上的预测结果 Y 应满足相等机会原则
- **审计模型:**
  - 审计数据生产者：查数据生产者生成的特征向量Z是否真正满足了脱敏要求
  - 审计数据用户：检查数据用户训练的最终模型 f 是否满足了预定的公平性标准

实现公平AI的三种策略

- **后处理 (Post-processing)：在模型训练完成之后对模型的输出进行处理，以满足公平性要求**
  - **方法：**为不同敏感属性群体设置不同的分类阈值或校准模型输出以确保不同群体间的决策统计量（如truth positive或false negative）达到平衡
  - **优点：**可以应用于任何已训练好的模型，无需修改模型的训练过程
- **预处理 (Pre-processing)：在模型训练开始之前对数据进行处理以去除偏差或提取不包含敏感信息的表示**
  - **方法：**偏差移除：如通过重采样、重新加权或数据转换等技术来改变训练数据的分布，使其对敏感属性脱敏；表示提取：通过学习一个特征表示 Z，使得 Z 中不再包含关于敏感属性 S 的信息
  - **优点：**训练出的模型天然地从公平的数据或特征中学习，有助于防止模型在训练过程中习得偏差
- **过程中处理 (In-processing)：在模型训练过程中直接将公平性要求纳入目标函数或优化过程中**
  - **方法：**施加约束：公平性指标作为一个正则化项或约束条件添加到模型的损失函数中；使用对抗网络：训练一个对抗网络，试图从模型的特征表示中预测出敏感属性 S，同时训练模型让对抗网络误差最大
  - **优点：**实现性能和公平性之间的最优权衡

今日课程内容

- ✓ ■什么是可信AI
- ✓ ■对AI的预测结果进行解释
- ✓ ■医学场景案例分析
- ✓ ■AI公平性定义
- ✓ ■公平AI的方法框架
- 👉 ■个体公平与群体公平

个体公平

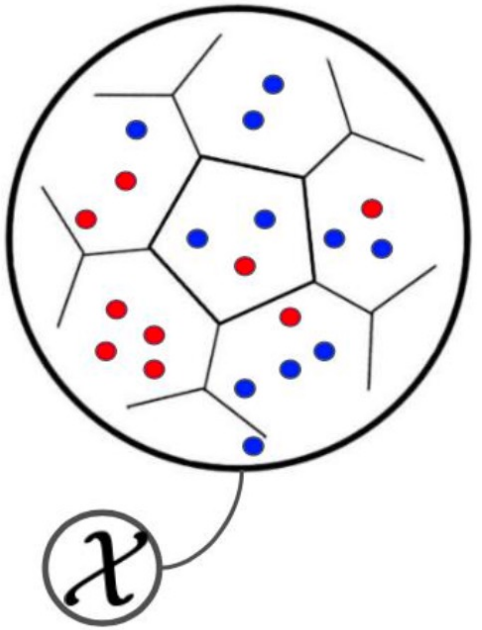
- **个体公平的核心原则：相似的个体应该被相似地对待**
  - 如果两个输入样本 x_i和 x_j 在某个相关指标上非常相似，那么模型对它们的预测输出 y_i 和 y_j 也应该非常相似，这要求我们定义一个合适的**距离度量** $\rho(x_i, x_j)$ 来衡量相似性
  - 如果两个在相似的个体（特征空间中距离很近）被分到了决策边界的两侧（即被不同分类），或者需要非常小的扰动就能改变一个样本的分类，就说明模型在该局部区域是**不鲁棒**且**不公平**的

问题：从事相同运动的相似个体对却被以不同的方式分类

- 模型对具有某些特征的个体存在偏差

数据管理者解决个体公平问题

- **要解决的根本性问题：**哪些个体是相似的？哪些个体应该被相似对待？
  - **需要定义公平性度量** $\rho(x_i, x_j)$（即在特征空间中的距离），并且需要基于领域知识、法律或伦理考量来确定，而不是简单地基于数据的原始特征
- **方法：**
  - **定义空间划分：**将输入空间 X 划分为**不相交的单元，要求**相似的个体位于同一个单元内
    - 如在信贷审批中，所有收入、信用历史和债务水平在特定范围内的申请者，无论其种族或性别，都应该被划分为同一个单元
  - **相似对待规则：**同一单元内的个体应该被相似对待，即使他们表面上看起来不同
    - 一旦个体被划分到同一个单元，意味着在**监管者定义的相似性标准下**他们是等价的，因此模型对他们的预测都应该保持高度一致



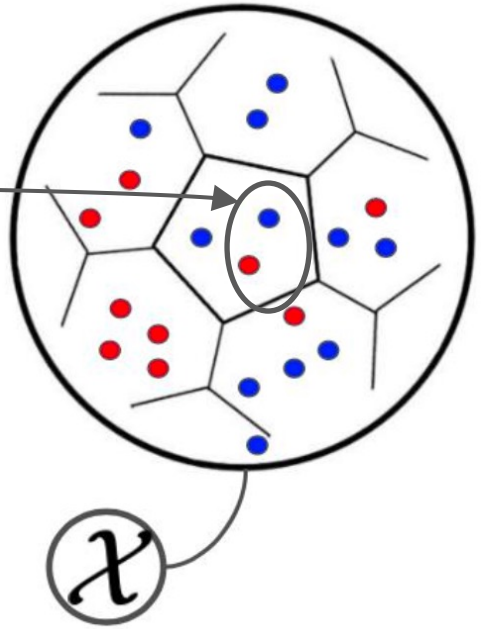
数据管理者解决个体公平问题

个体公平性蕴含着模型的鲁棒性

（对于输入中的微小扰动不敏感）

- 关键：提供了一个基于**分组和距离约束**的个体公平性定义
- 个体公平性定义：
  - 一个算法模型 A_D（依赖于数据 D 的算法）被称为$(B, \epsilon(D))$-个体公平，如果输入空间 χ可以被划分成 B 个不相交的子集（分组），记作 $\{C_i\}_{i=1}^{B}$，使得满足以下条件：

$$\forall x_1, x_2 \in \mathcal{X}, \quad \underbrace{(x_1, x_2 \in C_i)}_{\text{相似个体}} \Longrightarrow \underbrace{|l(A_\mathcal{D}, x_1) - l(A_\mathcal{D}, x_2)|}_{\text{算法对个体的预测损失}} \leq \underbrace{\epsilon(\mathcal{D})}_{\text{差异在容忍度内}}$$



如果两个个体被数据管理者定义为相似的，那么算法对它们的处理结果的差异必须在一个很小的容忍度内

个体公平性的优点与缺点

优点：

- **直观且易于解释：**“相似的个体应该被相似对待”原则符合人类的公平直觉，易于向数据生产者和非专业人士解释和传达
- **蕴含泛化能力：**模型在局部是平滑的（即相似输入得到相似输出）有助于模型在整个数据空间上的泛化能力，因为它减少了局部过度拟合或对微小变化的敏感性
- **蕴含统计平等：**个体公平蕴含着群体公平，通过专注于局部公平性，可以在宏观上提供群体公平性的保证

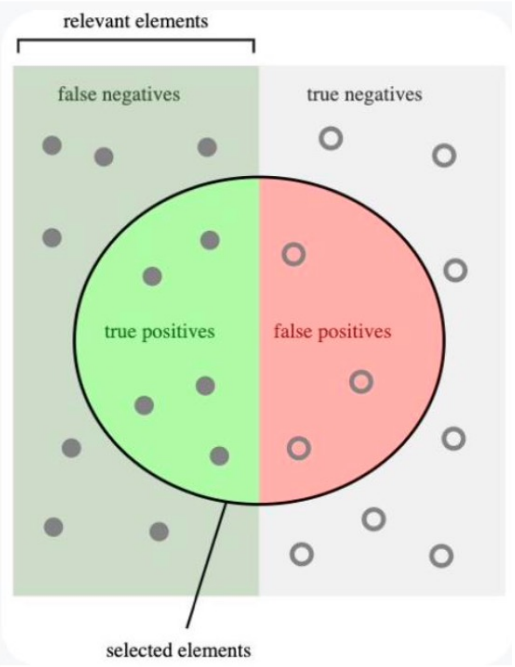
缺点：

- **监管者必须提供度量标准：**个体公平性的成功完全取决于数据监管者能否提供一个有效的相似性度量
- **构建度量标准需要专业知识：**需要深入的领域专业知识和人类洞察力以确定哪些特征是相关的或敏感的
- **公平性严重依赖度量标准的质量：**生成的“公平表示”的质量严重地依赖于监管者选择的相似性度量的质量
- **计算成本高昂：**通常涉及检查输入空间中大量相似对之间的距离和输出差异

群体公平

核心：关注在不同群体之间保持分类器统计数据的一致性

- **对于数据管理者的问题：**确定应该在跨敏感群体 S 上均衡哪个关于模型性能的统计量 $v(f, Y|S)$
- **典型的公平性统计量：**
  - 机会均等：True Positive Rate，$TP_S = P(Y = 1, f = 1|S)$
    - 如果真实结果为阳性，模型预测结果为阳性的概率在各群体间相等
  - 均衡赔率：True Positive Rate and False Positive Rate，$\{TP_S; FP_S\}$
    - 要求模型预测结果为真阳性和假阳性的概率在各群体都相等
  - 统计平等：Selection Rate，$TP_S + FP_S = P(f(Z) = 1|S)$
    - 要求模型对所有群体做出阳性预测的概率相等，与个体的真实结果 Y 无关



当在不同的敏感群体 S 中观察这张图时，TPs、FPs 或 (TPs+FPs) 所占的比例，必须在各个群体中保持一致

群体公平——统计平等（statistical parity）

1. 统计平等是一种流行的群体公平性度量：

- **基本元素：**
  - 总体：一个数据集 X，包含所有个体
  - 受保护子集：是总体 X 的一个子集 $S \subset X$，由一个敏感属性来定义
- **示例：**
  - X是待分类的人群样本
  - S是一个敏感或受保护的属性，如种族或性别
  - 统计平等要求：$P(Model\ Predicts \mid S) = P(Model\ Predicts \mid \bar{S})$
  - 含义：模型向受保护群体 S 做出积极预测的概率，必须与模型向其他群体 $\bar{S}$ 做出积极预测的概率相等
  - 关键点：只关注模型预测结果的分布，即所有群体被选中的比率一致

群体公平——统计平等（statistical parity）

2. 核心假设：在 X 上存在某种分布 D，它表示任何个体被选中进行评估的概率

- 分布 D 描述了在实际应用中遇到不同个体的频率
- 在机器学习中，通常假设训练和测试数据是从这个真实的、潜在的总体分布 D 中独立同分布 (i.i.d.) 抽样得到的
- 实际的评估分布 D 反映了那些真正寻求评估（如是否得到重点医疗资源）的人群的分布
- **一般性原则**：对 D 不施加任何限制，公平性的定义将适用于任何 D
  - 无论实际人群的分布 D 是什么样子（无论多少人真正申请医疗资源），只要模型 f 的输出满足该条件，它就被认为是统计平等的
- **意义**：将公平性定义与实际人群的抽样分布分离开来
  - **对于任何被评估的个体，模型在不同敏感群体上做出的“积极”或“被选中”的预测比例是否一致，而不去担心为什么有些人一开始就没有参与评估**

群体公平——统计平等（statistical parity）

3. 统计平等的形式化表达与不利影响

- 分类器: $f: X \to \{0,1\}$
- 统计不平等：衡量模型 f 在敏感属性 S 上的**不公平程度**的指标：

$$\text{imparity}_f(X, S, D) = P(f(x) = 1 \mid x \in S^c) - P(f(x) = 1 \mid x \in S)$$

从非受保护群体中随机抽取的个体被预测为1的概率　**从受保护群体**中随机抽取的个体被预测为1的概率

- 如果模型是完全统计平等的，那么这个**不公平程度**应该等于 **0**
- **不利影响**：衡量多数群体和受保护群体获得特定结果的差异
- 如果不公平程度是正数且较大，表示非受保护群体 S^c 被预测为 1 的概率显著高于受保护群体 S
  - 该差异就为**不利影响**的体现，意味着受保护群体在获得积极结果方面受到了不利待遇

群体公平——统计平等（statistical parity）

4. 统计平等的形式化总结

- **统计不平等与统计平等的关系：**
  - 统计不平等衡量了多数群体和受保护群体获得特定结果的差异
  - 当这种差异很小，分类器就具备统计平等，即符合公平性的概念
- **正式定义：**
  - 模型 f 在分布 D 上对敏感属性 S 的不公平程度小于偏差 ϵ，即：

$$|\text{imparity}_f(X, S, D)| < \epsilon$$

$$|P(f(x) = 1 \mid x \in S^c) - P(f(x) = 1 \mid x \in S)| < \epsilon$$

  就说明模型对敏感属性达到了统计平等
- **实际意义：**统计上相似地对待普通总体和受保护群体
- **潜在问题：**效率悖论，即如果受保护群体的实际资质低于非受保护群体，强制统计平等可能会损害模型的准确性或社会目的

群体公平策略的优点与挑战

优点：

- 计算、度量和执行效率高：只需要计算几个群体的汇总统计量，计算成本通常比个体公平性低得多
- 通常更容易向政策制定者解释：群体公平性是宏观和社会性的，更容易被政策制定者和非专业人士理解

挑战：

- 监管者必须确定要均衡的分类器统计量：存在多种群体公平性度量（如统计平等、机会均等、均衡赔率），它们通常**彼此冲突**，需根据领域和伦理要求，选择最合适的度量
- 公平性依赖于监管者所选度量标准的质量：不恰当的度量标准导致模型将**牺牲准确性**来满足这个可能带有误导性的公平性要求
- 群体公平性可能导致违反个体公平性：群体公平性是**粗粒度的方法**，即使群体满足公平性，群体内部的个体差异仍可能被忽略。此外，个体可能同时属于多个敏感群体，存在**多维度交叉问题**
- 可能导致公平性操纵：模型可以通过**有目的地利用未受保护的特征**来满足表面上的群体公平性统计量

See you next week !