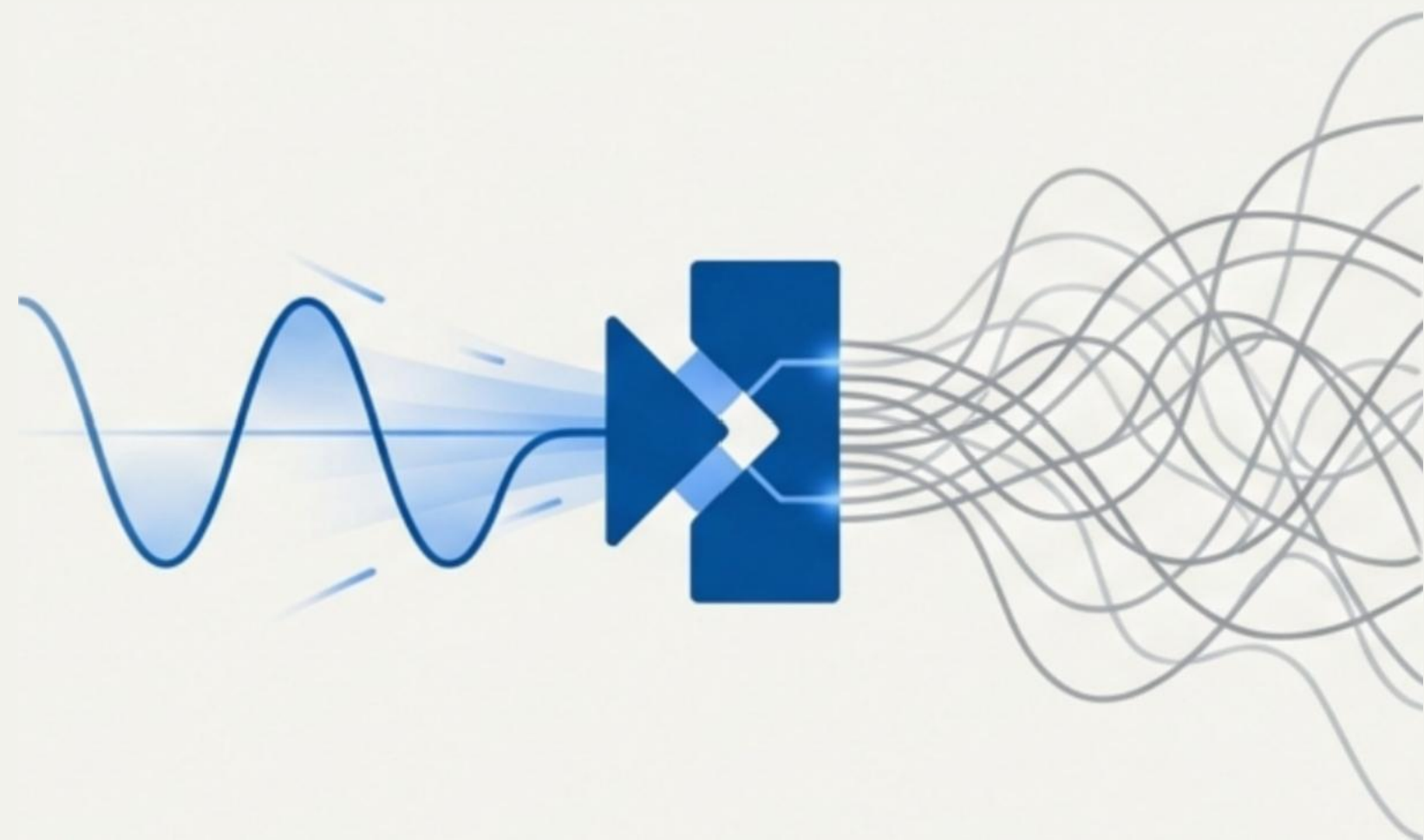


# OpenRisk-Advisor: 开源项目衰退风险预测与智能治理建议系统

本项目基于大赛官方推荐的开源项目与工具集构建，包括 OpenDigger, IoTDB, DataEase 等。  
秉持拥抱AI的理念，本PPT借助AI进行部分绘图，旨在阐述设计理念，并非产品demo。

队名：大东北我的家乡  
学校：华东师范大学  
成员：任北鸣张凌琨



# 背景与挑战：开源项目的困境

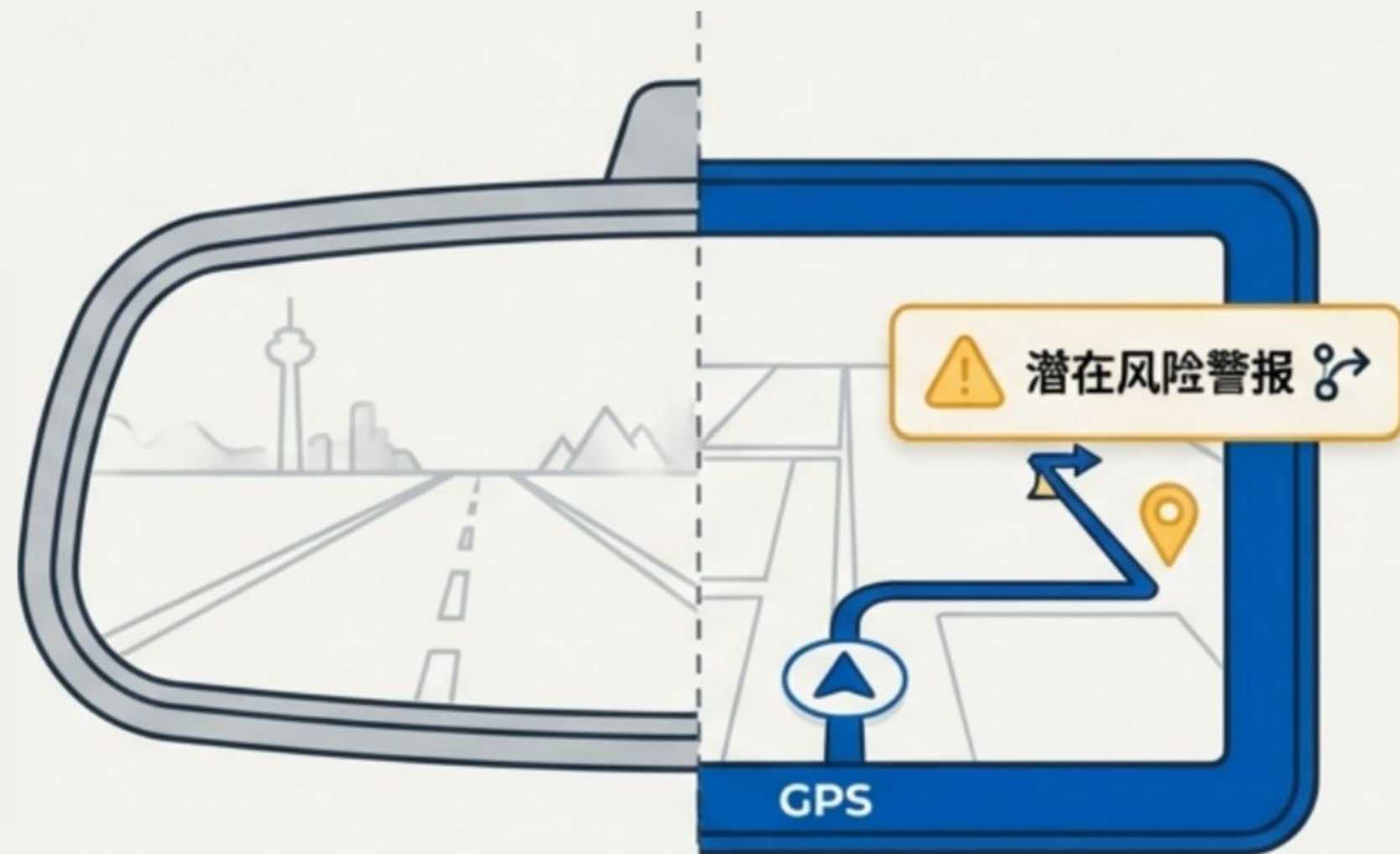
缓慢衰退比突然死亡更常见，也更难发现。

开源项目并非突然停止，而是经历一个缓慢的衰退过程：贡献者流失、协作效率下降、社区热度降低。当这些迹象变得明显时，往往已错过最佳干预时机。

现有工具的局限性：“后视镜”而非“导航仪”

- 描述历史，而非预测未来：大多工具停留在历史数据展示与静态评分，无法提供前瞻性预警。
- 信号与噪声混杂：难以区分“正常周期波动”（如节假日、版本发布节奏）与“异常持续下滑”（真正的衰退信号）。
- 缺乏可行动的建议：即便发现问题，也无法提供“如何介入”的具体方案。

本项目核心挑战：将生态数据转化为可信的、可行的、面向未来的治理决策依据。






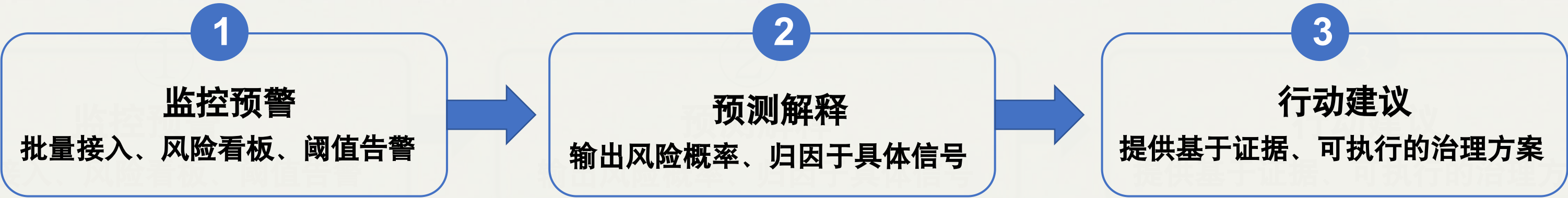
历史数据

风险预测与建议

# 目标用户与核心需求：从数据洞察到决策支持

三类核心用户， 一个共同目标： 提前发现并有效干预项目风险。

用户画像	核心痛点	系统价值
 开源项目维护者	“我的项目在走下坡路吗？” “我该从何处着手改进？”	快速定位健康度瓶颈 获取具体、可执行的治理建议
 企业OSPO/ 技术管理者	“我们依赖的几百个项目哪个有风险？” “如何为技术选型提供依据？”	低成本监控依赖风险 为资源投入与替代方案提供数据支持
 开源基金会/社区运营	“哪些项目最需要扶持？” “如何量化评估运营活动效果？”	精准识别潜力与风险项目 数据驱动的资源配置与效果评估

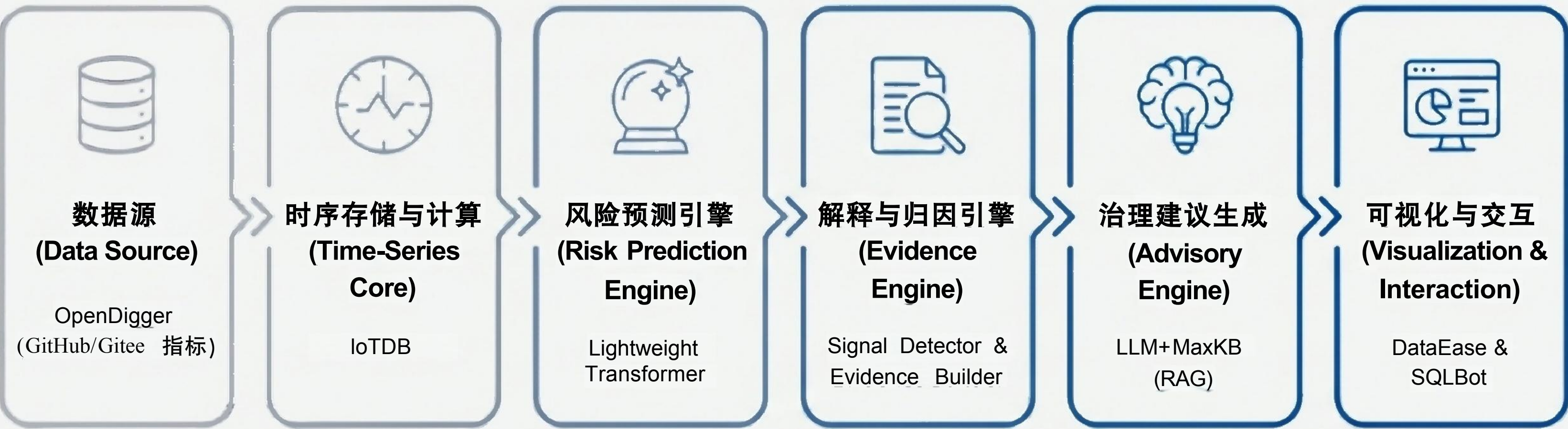




# 总体架构：一条从“数据”到“决策”的智能流水线

设计理念”：分层解耦、稳定接口、可迭代升级。

我们将开源治理问题建模为一条标准化的智能处理流水线，确保各模块职责清晰，系统可复现、可审计。



核心优势： 将“黑盒”洞察转变为“白盒”决策支持，每一步都可追溯。

# 数据与信号：构建可复现、可审计的治理“词典”

## 数据基础：OpenDigger官方口径。

- **数据源**：OpenDigger 静态指标数据根 (JSON 文件)
- **时间粒度**：以月为核心预警单位，季度/年度为长期参照
- **指标体系**：围绕“可持续协作”构建，覆盖活跃度、贡献者、协作效率、社区互动四大维度。

## 核心方法：从“原始指标”到“治理信号”的转化

- 我们将模糊的“衰退感觉”分解为一组可配置、可审计的结构化信号。

**可复现承诺**：通过 `metrics.yaml` 和 `signals.yaml` 配置文件，保证数据口径与信号规则的完全可复现。



# 时序底座：loTDB

开源生态数据是典型的“多实体 x 多测点”时序场景，使用通用数据库(如 MySQL)可能会在数据对齐、窗口计算和查询性能上付出巨大工程代价。

## loTDB 的三大核心优势，完美匹配本场景

特性	通用数据库(痛点)	Apache loTDB(解决方案)
数据对齐	ETL逻辑复杂，跨指标JOIN成本高	<ul style="list-style-type: none"><li>原生支持时间序列对齐(ALIGN BY TIME)</li><li>查询侧解决对齐问题</li></ul>
窗口聚合	需手写复杂SQL,难以复用	<ul style="list-style-type: none"><li>内置高效的窗口聚合与下采样函数(GROUP BY)</li><li>计算下推至存储层</li></ul>
缺失值处理	每个应用需单独处理NULL值，逻辑分散	<ul style="list-style-type: none"><li>查询侧提供多种填充策略(FILL)</li><li>保证下游输出稳定</li></ul>

## 我们的数据模型



实现原始序列与衍生特征分层存储，保证数据可追溯、特征可复现。

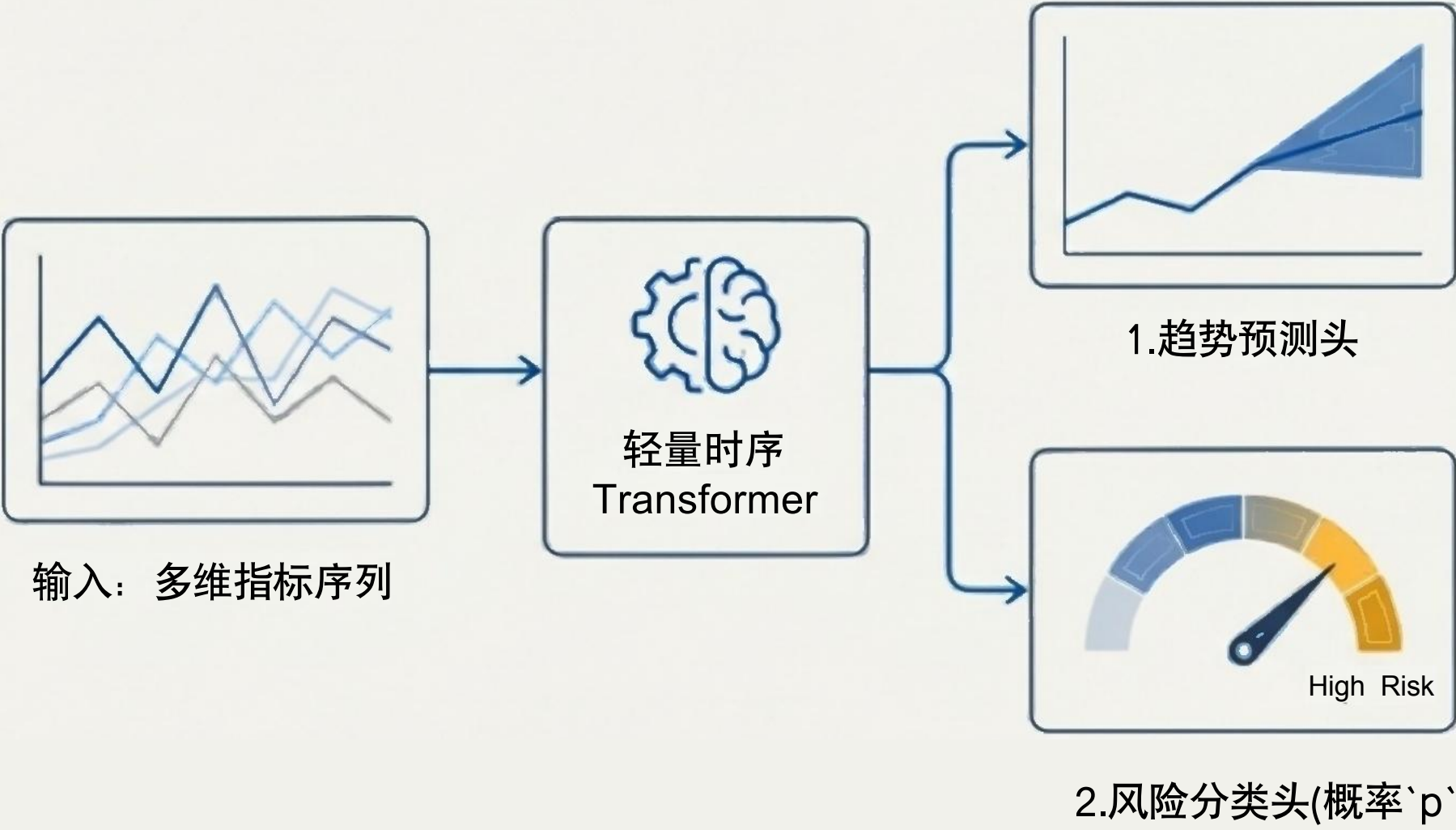


# 风险预测：从历史数据到未来趋势的概率化洞察

**任务定义：** 一个弱监督、多任务的时序学习问题。在缺少人工标注的情况下，我们通过规则自动生成弱监督标签，并预测未来风险。

## 模型方案：轻量级时序Transformer+ 双输出头

- **输入：** 过去12-18个月的项目多维指标序列(月度)
- **主干：** 轻量Transformer Encoder,捕捉跨指标、长周期的复杂依赖关系
- **双任务输出：**
  - 1. **趋势预测头：** 预测未来1-3个月关键指标的趋势与不确定性区间
  - 2. **风险分类头：** 输出未来窗口内发生“持续性衰退”的概率`p`



## 创新亮点

- **弱监督学习：** 无需人工标注，系统可持续自迭代。
- **多任务学习：** 趋势预测为风险判断提供证据，使训练更稳定，结果更可信。
- **概率化输出：** 风险不再是“是/否”，而是可校准、可配置阈值的概率，更贴合真实决策场景。

# 解释与归因：构建可行动的证据链

**设计原则：** 解释的目的是为了行动，而非学术指标。我们拒绝“黑盒”分数。系统必须回答：风险从何时开始?有多严重?意味着什么?

## 结构化的风险报告(RiskReport)

解释引擎将所有分析结果打包成一个标准化的JSON对象，作为系统内所有后续模块(LLM、可视化)的唯一事实来源。

```
{
  "risk_score":0.85,
  "risk_level":"High",
  "main_signals":[
    {
      "signal_name":"协作效率持续恶化",
      "strength":0.9,
      "time_window":"2024-01 to 2024-83",
      "supporting_metrics":{"
        "issue_age_avg":"+35%vs prev_window" }
    },
    "model_uncertainty":"Low"
  ],
  ...
}
```

价值： 将模型的模蝴输出， 转化为可验证、可复现、可审计的结构化证据。



# LLM 治理建议：可控的、基于证据的行动生成器

**严格定位：** LLM 是“行动建议助手”，而非“风险决策者”。为了根除幻觉、保证工程可信度，LLM 不直接接触原始数据，其唯一的输入是上一页生成的RiskReport 证据对象。

## 工作流程：检索增强生成 (RAG)+ 模板化输出

Step 1:输入 (Input)

```
{
  "risk_score": 0.85,
  "risk_level": "High",
  "main_signals": [
    {
      "signal_name": "协作效率持续恶化",
      "strength": 0.9,
      "time_window": "2024-01 to 2024-03",
      "supporting_metrics": {
        "issue_age_avg": "+35% vs prev_window"
      }
    }
  ],
  "model_uncertainty": "Low"
}
```

Step 2:检索 (Retrieve)



根据 main\_signals 从知识库中检索相关的治理最佳实践(如 Issue 分流策略、贡献者激励机制)。

Step 3:生成 (Generate)



LLM 结合证据和知识，填充一个严格的模板，生成包含“目标、步骤、预期效果”的行动计划。



### 安全约束

- ✓ 证据驱动：所有建议必须明确关联到 RiskReport 中的一个或多个信号。
- ✓ 知识可溯：生成内容必须引用 MaxKB 中的知识来源。
- ✓ 杜绝新事实：禁止生成 RiskReport 中未包含的任何数据断言

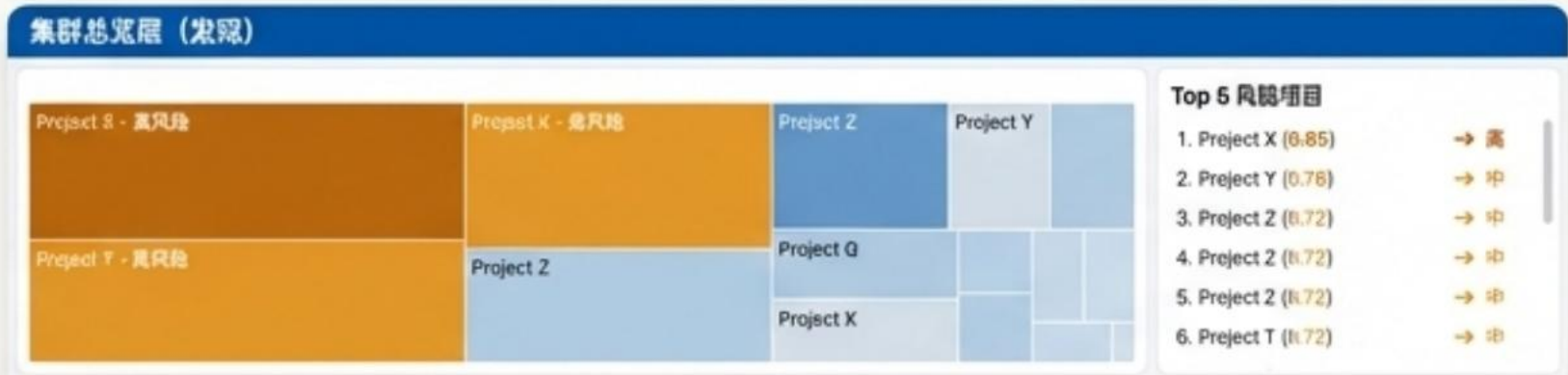
# 可视化与交互：从“发现”到“行动”的闭环体验

**工具选型：** DataEase 实现快速原型与高效交付。我们利用构建了服务于不同决策层级的交互式看板。

## 1. 集群总览层(发现)

视图：风险地图、TopN风险项目榜单

目标：1分钟内定位哪个项目最需要关注。



## 2. 单项目诊断层(诊断)

视图：关键指标时序图、信号触发区间标注、证据摘要

目标：验证风险、定位具体恶化维度。




## 3. 行动建议层(行动)

视图：LLM生成的治理建议、关联证据与知识来源

目标：提供可执行、可跟踪的改进方案。



 **交互辅助：** SQLBot 提供自然语言查询入口，赋能临时性、探索性的数据分析。



# 工程计划与里程碑：确保方案的可落地性

模块化开发：四大后端服务+一个展示层，并行推进。



**数据服务：** 对接OpenDigger，实现数据标准化入库



**特征服务：** 基于IoTDB，提供统一时序特征计算与查询



**风险服务：** 运行模型与规则，生成RiskReport



**建议服务：** 调用LLM 与 MaxKB，生成治理建议



**展示层：** DataEase看板与SQLBot查询

## 里程碑规划(初赛 → 复赛)

初赛

复赛

M1-核心链路验证

完成数据采集、IoTDB 存储、基线风险规则与初步看板搭建，跑通端到端闭环。

M2- 智能模型集成

部署时序Transformer模型，输出概率化风险与趋势预测，升级RiskReport证据链。

M3- 建议系统上线

对接LLM与 MaxKB，实现基于证据的建议生成与知识追溯。

M4- 交付与展示

完善部署文档，提供可公开访问的演示环境，完成作品材料。



# 创新性与差异化：从“事后归因”到“事前干预”

将开源治理从“静态描述”升级为“动态预警与智能干预”。

对比维度	传统生态分析工具	OpenRisk-Advisor（本项目）
分析视角	历史描述： 静态健康度评分、历史趋势图 	未来预测与建议 衰退风险概率、未来趋势预测、行动建议 
解释性	数据罗列： 用户自行解读图表 	结构化证据链自动关联“信号-指标-时间”，输出可机读的RiskReport 
AI应用	黑盒模型： 难以信任和追溯 	可控的白盒AI LLM仅作为“证据翻译器”，流程可审计 
系统闭环	信息展示： 止步于“看” 	决策闭环 实现“预警→解释→建议→评估”的完整治理流程 

我们并非创造一个新算法，而是面向真实场景，创造性地组合现有工具，构建了一套工程上可控、价值上闭环的治理系统。

# 应用价值与推广路径： 赋能开源生态可持续发展

## 对生态的直接价值



### 赋能维护者

从被动响应到主动治理，降低心智负担。



### 赋能 OSPO

从人工排查到智能扫描，高效管理开源依赖风险。



### 赋能基金会

从经验驱动到数据驱动，精准配置扶持资源。

## 推广与落地路径

1

### 开源交付

项目以开源形式发布，提供一键部署脚本与演示数据，降低使用门槛。



2

### 试点合作

与头部开源社区或企业 OSPO 合作，进行真实场景验证与迭代。



3

### 平台化扩展

将能力抽象为可复用的治理平台，未来可扩展至安全漏洞、版本节奏等更多风险场景。

# 总结: OpenRisk-Advisor, 开源生态的智能风险“哨兵”

## 我们实现了



定义了一个真实问题: 开源项目的“缓慢衰退”难以察觉。



构建了一套闭环方案: 实现了从“风险预测”到“证据归因”再到“行动建议”的全链路。



展示了工程可控性: 通过结构化证据链约束LLM, 确保了系统的可信与可复现。



探索了工具集价值: 实践了OpenDigger、IoTDB等开源组件在协同工作下的巨大潜力。

## 我们的愿景

我们致力于将OpenRisk-Advisor打造成开源社区与OSPO的标准化基础设施, 让每一个开源项目都能拥有自己的“健康顾问”。