

科研素养训练II

第2次课：复现目标界定与方案设计

宋庆禹
2026.09.14



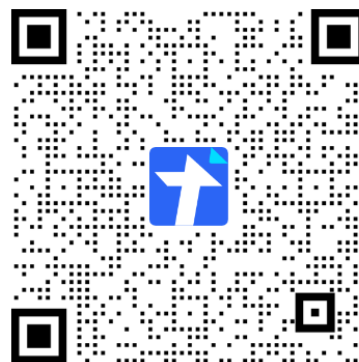
论文复现组队+选论文

- 要求：4人一组
- 截止时间：9.30日23:59



2026Fall科研素养训练II...

扫一扫二维码打开或分享给好友



- 腾讯文档 -

可多人实时在线编辑，权限安全可控



本次课内容

- 1 选择复现目标
- 2 界定复现问题
- 3 实验方案的结构化
- 4 成功判别依据与方案审查

课堂产出

论文复现方案书初稿

本课程提供复现方案的基本检查框架。
可以提出不同方案，但应说明选择理由、适用条件和局限。



本次课程的学习任务

本次课回答两个问题：准备复现什么？准备怎样复现？

输入

- 第1次课完成的论文阅读卡
- 论文中的核心观点与对应证据的位置
- 研究问题、方法、数据和指标

输出

- 一个范围明确的复现目标
- 一套可以执行的实验方案
- 预先确定的评价指标与判据



课堂问题：有限条件下如何选择

如果只有一周时间和一台普通计算机，你会选择什么？

A 论文中的全部实验

B 摘要中的主要结论

C 一张关键图或一个关键表

D 最容易运行的一段代码



课堂问题：有限条件下如何选择

如果只有一周时间和一台普通计算机，你会选择什么？

A 论文中的全部实验

B 摘要中的主要结论

C 一张关键图或一个关键表

D 最容易运行的一段代码

条件判断：C一定程度兼顾核心意义、可验证性和可行性



复现对象的范围

范围模糊

- 复现这篇论文
- 验证方法M是否有效
- 把作者代码运行起来

范围明确

- 定位到论文图5
- 比较方法M与基线B
- 使用数据集D和宏平均F1

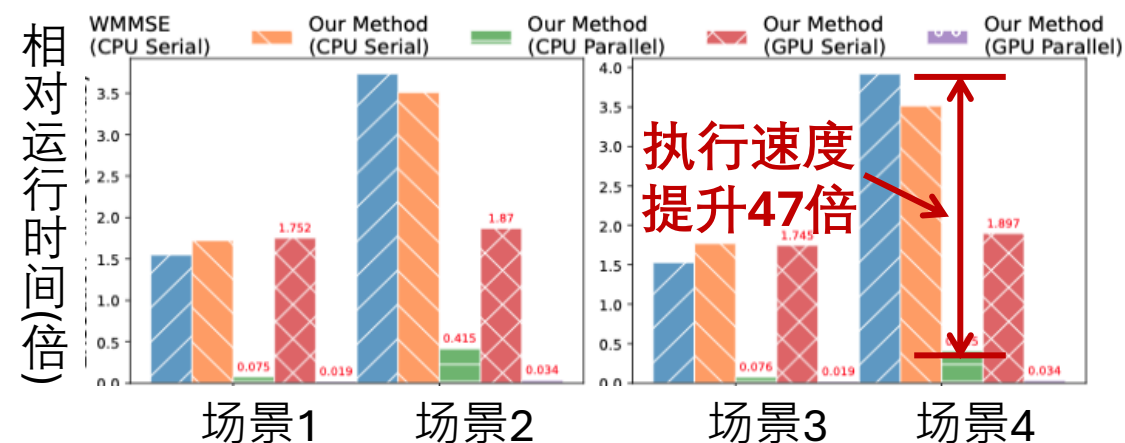
复现目标应能定位到具体图表、指标和实验条件

示例：Stanford CS244的项目选择方法

Stanford CS244的项目起点

- 学生两人一组开展论文复现实验
- 复现目标：从论文中选择一张关键图或一个关键表
- 目标必须关联论文的研究动机或核心观点

论文实验结果示例



来源: Yan & McKeown, Stanford CS244



论文中的可复现对象

关键数值

准确率达到92.4%

相对提升

比基线提高4个百分点

变化趋势

规模增加时性能保持稳定

方法排序

方法M优于B和C

系统性质

高负载下保持低时延

消融结论

模块X贡献最大

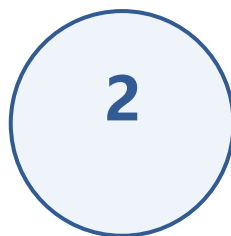
优先选择能够直接支撑论文核心结论的结果



复现对象的五项选择标准



重要性



可验证性



资源可得



结果可比



(课程) 可完成

一个目标只有在科研意义和现实条件之间取得平衡，才适合作为复现任务



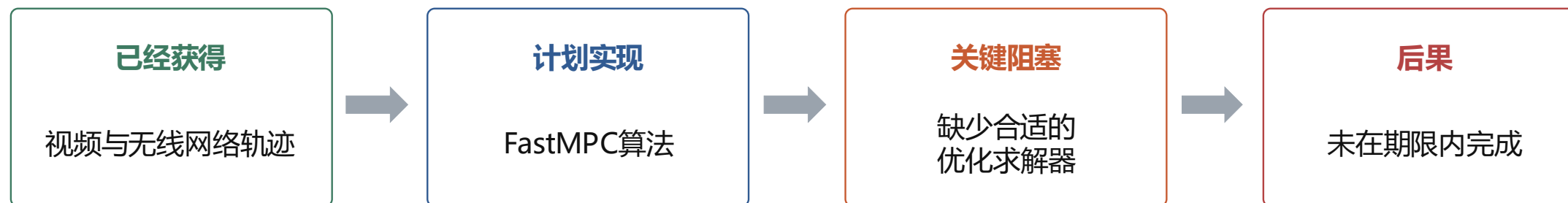
示例：候选目标评估

候选目标	重要性	可行性	判断
完整复现大规模集群实验	5	1	范围过大
复现核心算法在公开数据集上的结果	5	4	优先选择
运行作者提供的演示程序	2	5	仅适合作为环境测试

优先选择核心、有限、可比较的结果



案例：高估带来的项目风险



MPC算法：模型预测控制（Model Predictive Control），示例：

$$\min_{\{u_k\}} \sum_{k=0}^{N-1} (x_k^\top Q x_k + u_k^\top R u_k) + x_N^\top P x_N \quad x_{k+1} = A x_k + B u_k,$$

项目受阻往往源于未被提前识别的关键工具，而非不努力

来源：Stanford CS244 FastMPC案例：

<https://reproducingnetworkresearch.wordpress.com/2016/05/30/cs244-16-failed-experiments-with-fastmpc-integrating-rate-based-adaptive-streaming-into-vlc/>



最小可复现主张 (Claim)

本课程使用“最小可复现主张”控制项目范围

范围过大

复现方法M的全部实验

范围适当

复现M在数据集D上相对基线B的F1提升

范围过小

验证某代码段可以打印出一个数值

在不破坏论文核心逻辑的前提下，可以独立验证的最小结果单元

$$F1 \text{ Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

		Prediction	
		Positive	Negative
Actual	True	TP	FN
	False	FP	TN

$$\text{precision} = \frac{TP}{TP + FP}$$

$$\text{recall} = \frac{TP}{TP + FN}$$

第一位T/F: 表示预测的对错
第二位P/N: 表示预测的结果

TP: True Positive, False Negative

$$F1 \text{ score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$



复现拆解：研究内容的结构化拆解

“方法M在数据集D上的F1达到0.842，比基线B提高4.1个百分点。”

组成部分	从主张中提取的信息
对象	方法M
条件	数据集D及论文评价协议
比较	基线B
指标	宏平均F1
预期结果	0.842，比基线提高约4.1个百分点



复现问题构建：从论文陈述到复现问题

论文中的陈述

- 方法M优于基线B
- 方法M具有更好的性能
- 方法M具有较强稳定性

复现中的问题

- 在相同协议下，M是否高于B？
- 提升幅度是否接近原论文？
- 不同随机种子下排序是否保持？

陈述描述作者得到的结果，问题规定需要收集的证据



复现问题的必要信息

一个完整的复现问题包含五项信息

对象

条件

比较

指标

判断

方法M

数据集D与固定划分

基线B

平均F1

是否仍然更高

在数据集D与固定划分下，方法M的平均F1是否仍高于基线B？



复现问题的可检验性

不可检验的表述

- 方法M的效果好吗?
- 结果是否比较接近?
- 模型是否足够稳定?

可以检验的表述

- M的F1是否高于B?
- 与原结果相差是否不超过0.02?
- 5次运行中M是否至少4次优于B?

可检验意味着不同结果能够导向不同判断



复现问题的三类可能结论

支持

M高于B，差值落入预设范围

不支持

M未高于B，或方法排序发生
反转

无法判断

数据、指标或运行条件无法对
应

方案应当提前说明三类判断分别对应什么证据



复现方案中的透明记录

开放、可信的方案需要区分三类信息

论文明确说明



数据集版本、参数、评价指标

自行假设



缺失参数的取值、替代工具

目前尚未确认



代码对应版本、预处理细节

资料来源: [LMU与FORRT开放研究教学资源](#)



课堂练习：界定一个可检验的复现目标

案例材料 (Toy Example)

3分钟

某论文表2显示：

方法M的宏平均F1为0.836，基线B为0.802。作者认为，M在该实验条件下优于B。

请用一句话回答

**如果要验证这一比较结论，你准备做什么实验？
本次明确不做什么？**

示例：找到原文对应数据集D，按照论文规定的划分、预处理和评价条件，分别运行方法M与基线B，计算并比较两者的宏平均F1，检验M是否优于B；本次不扩展到其他数据集，也不尝试改进方法M。

提示：说清比较对象、实验条件和评价指标，并限定范围。



第一学时检查点

第一学时检查点

我们准备在条件 C 下，使用指标 I，复现论文结果 R，并与基线 B 比较

- ☐ 能够定位到具体证据
- ☐ 可以通过实验回答
- ☐ 具有核心研究意义
- ☐ 能在课程条件下完成



同一篇论文的不同复现路径

直接运行

使用作者代码、数据与配置

重新实现

依据论文描述自行实现

替代平台

在其他框架或硬件上重建

条件检验

改变数据、参数或规模

路径不同，能够支持的结论也不同



复现路径与结论边界

复现路径	可以回答	不能直接回答
运行作者代码	公开材料能否产生相近结果	独立实现是否可行
自行重新实现	论文描述是否足以支持实现	作者代码是否正确
更换实验平台	结论能否跨平台保持	原平台上的精确结果
更换数据	结论是否具有一定泛化性	原始数据上的复现程度

做了什么，决定了最终能够说什么



案例：实验平台的选择

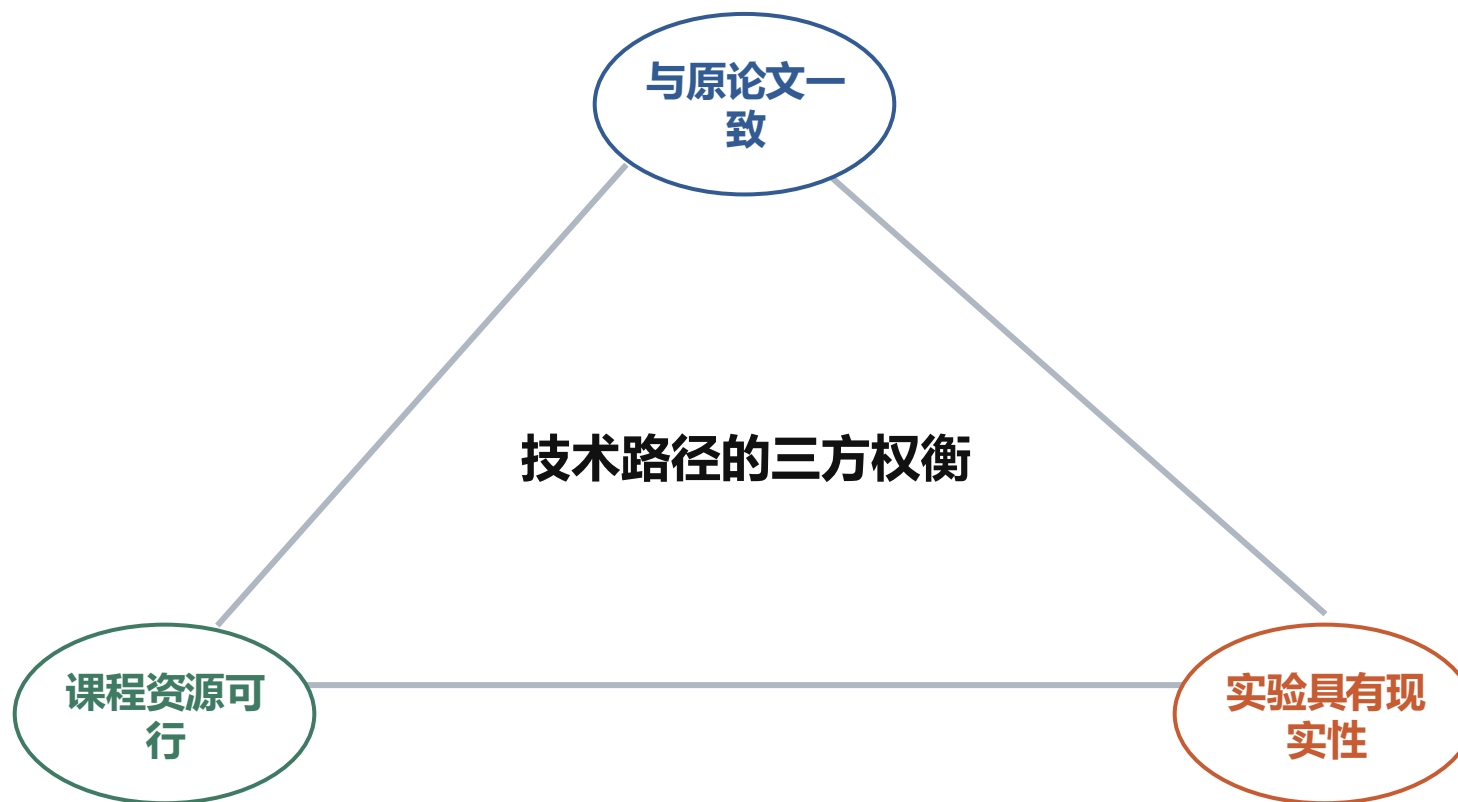
平台	适用实验	主要特征
Mininet	多节点网络拓扑	实时处理流量，便于构建网络系统
Mahimahi	单链路拥塞控制	适合链路条件与协议实验
ns-3	更大规模的模拟	当实验规模超出仿真环境能力时使用

平台应与研究问题、实验规模和比较目标相匹配

资料来源：[Stanford CS244平台选择案例](#)

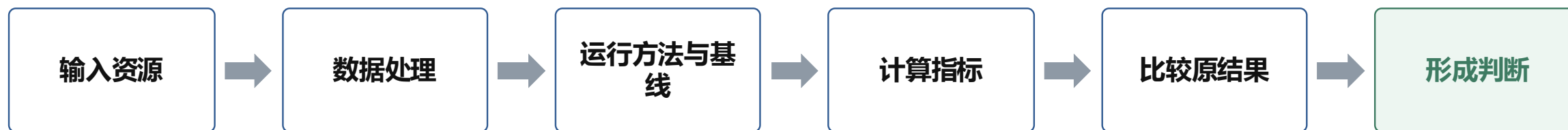


技术路径的三方权衡





复现实验的基本结构



实验方案需要明确每一步的输入、操作和输出



实验中的四类变量

自变量

主动改变或比较的因素

方法M与基线B

因变量

用于观察结果的指标

宏平均F1

控制变量

比较时保持一致的条件

数据划分与训练轮数

干扰因素

难以完全控制但可能影响结果

硬件与随机性



变量识别示例

变量类型	贯穿案例中的具体内容
自变量	使用方法M还是基线B
因变量	固定测试集上的宏平均F1
控制变量	数据版本、划分、预处理和训练轮数
潜在干扰因素	随机初始化、软件版本和硬件差异

第3次课将进一步处理资源、版本和环境问题



实验条件矩阵

实验编号	方法	数据	参数	随机种子	输出
E1	基线B	D	论文设置	1	F1
E2	方法M	D	论文设置	1	F1
E3	基线B	D	论文设置	2	F1
E4	方法M	D	论文设置	2	F1

条件矩阵帮助我们发现漏做的实验和不公平的比较



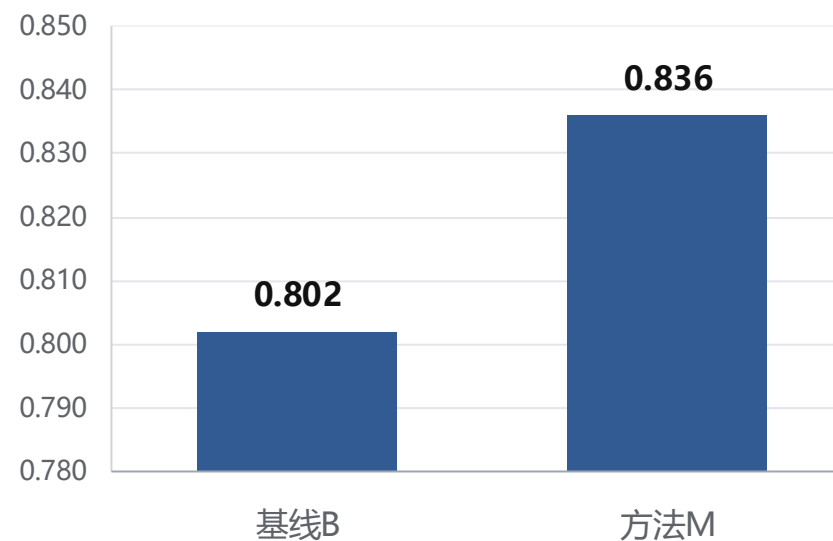
基线提供结果的比较背景

只运行方法M

F1 = 0.836

这个结果好不好？

同时运行基线B



示例数据，仅用于说明基线比较



实验协议示例：方法M与基线B的比较

复现目标：在论文指定的数据与评价条件下，检验M的宏平均F1是否高于B。

- 1 使用论文指定的数据版本及训练、验证、测试划分
- 2 按论文规定完成预处理记录具体操作
- 3 分别运行方法M与基线B记录配置，说明比较条件
- 4 **本示例：每种方法运行5次记录每次使用的随机种子**
- 5 在固定测试集上评价采用相同的宏平均F1计算方式
- 6 报告全部运行结果汇总均值、标准差和方法差值

重复运行用于观察随机性带来的波动。5次是本示例的安排，实际次数应结合原论文要求、预计波动与计算预算确定。

基本要求：他人能够按记录执行实验，并理解你的选择依据。



Stanford CS244的成功标准



重建实验

实验系统能够运行并产生结果



比较形态

复现结果与原结果的
趋势或形态相近



解释差异

能够说明差异的来源与影响

资料来源: [Stanford CS244的三项成功标准](#)



复现结果的三个比较层次

数值一致

结果是否落入预设容差

趋势一致

曲线形态、变化方向或排序是否一致

结论一致

差异是否改变论文的核心主张



复现成功与否的判定依据

模糊判据

- 结果大致一致
- 模型比较稳定
- 性能没有明显下降

可以操作的判据

- 与原结果相差不超过0.02
- 5次运行中至少4次高于B
- 相对提升方向与论文一致

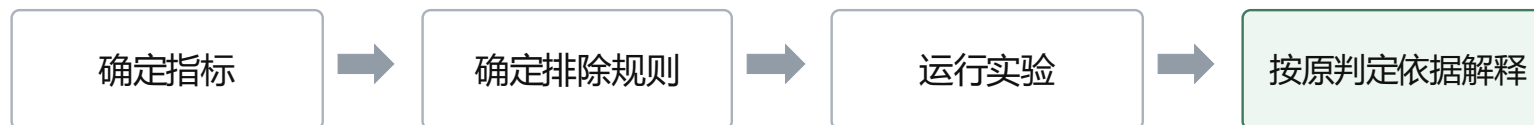
判断依据应当允许不同学生对同一结果作出相同判断



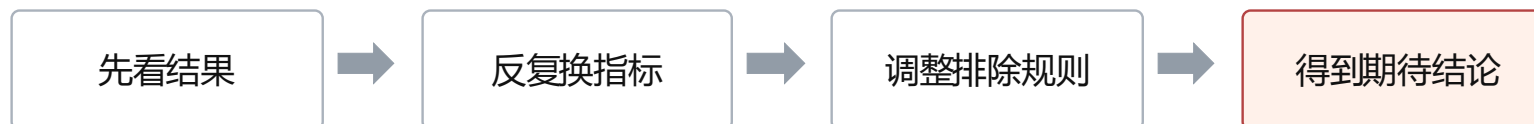
预先确定分析决策

先写判据，再查看结果

路径A



路径B



预先决策帮助区分计划内分析和看到结果后的调整

轻量预注册式复现方案

复现方案书可以视为一份轻量预注册式计划

待复现结果

实验前填写

实验条件

实验前填写

主要指标

实验前填写

重复次数

实验前填写

排除规则

实验前填写

成功判据

实验前填写

资料来源: [MIT Tools for Robust Science: 预注册分析计划](#)



复现风险清单

风险	可能影响	应对方案
作者代码不可用	无法直接运行	根据论文实现或缩小范围（需考虑论文描述是否足以支持实现）
数据版本不明	结果不可直接比较	查找归档版本并记录差异
计算资源不足	无法达到原实验规模	按比例缩小并保留关键关系（需确保保留要检验的关键条件）
关键参数缺失	结果波动较大	查证、作出明确假设、有限敏感性分析（需确认该参数是否可能决定主要结果）
工作量估计过大	无法按期完成	优先保证最小可复现主张



同伴审查复现方案

同伴审查：一个优点、一个风险、一个建议

- | | |
|--------------------------------------|---------------------------------------|
| ☐ 目标能否定位到具体结果 | ☐ 变量与控制条件是否明确 |
| ☐ 问题能否由实验回答 | ☐ 成功判定依据是否可以操作 |
| ☐ 基线与实验组是否公平 | ☐ 工作量是否能按时完成 |

2分钟阅读，2分钟填写反馈，1分钟交换意见

资料来源：[FORRT学生参与与开放研究教学实践](#)



作业2：提交论文复现方案书

- 待复现的图、表或核心结论
- 可检验的复现问题与范围边界
- 复现路径、实验流程与条件矩阵
- 评价指标、重复次数与成功判定依据
- 风险清单及备选方案

提交形式

复现方案书.pdf



本次课程小结与课后任务

复现目标



定位到具体图表与核心主张

实验方案



明确路径、变量、基线与步骤

判断规则



实验前写下指标和成功判据

下次课：复现资源管理与实验环境构建