

如何做研究-解决问题

浙江大学 陈为

目录

- 什么是解决方案
- 提出观点
- 给出验证
- 一些误区

1 什么是解决方案

解决方案

- 它是可行的，而且可以在合理的时间内完成。
- 执行它的代价会比没有解决这个难题的损失更低。
- 它不会制造比原来更大的难题。
- 它比其他可能的替代方案更经济与快速
 - 这可能是个非常难以论证的观点。

科研解决方案的特点

- Novel
- Nontrivial
- 相比已有方法有优点或改进

解决方案的贡献

- 最好的贡献：创新
 - 第一次提出人脸识别算法：天才的想法！
 - 第二次提出人脸识别算法：显然不是一个新颖的问题
 - 第一次提出一个蛇识别算法：够新颖，但应用价值不足

解决方案的贡献

- 次好的贡献：优化
 - 更快的速度
 - 更高的准确率
 - 更好的质量
 - 在理论上更完备
 - 提供了新的见解

提出优化

- 秘诀1：用A领域的方法解决B领域的问题
 - 改进了B领域问题的解决方法
 - 对B领域的问题提供了新的理解
 - 拓展了A领域方法的适用范围

提出优化

- 秘诀2：负面的结果也可能是有贡献的
 - 通过意外的负面结果，揭示未发现的机理
 - 展示已有工作的反例或错误

如：Heawood在1889年发表了一篇论文，指出Kempe在1879年关于四色定理的论断的一个反例

解决方案的贡献

- 较弱的贡献：组合
 - 将已有技术通过某种新颖的方式组合起来，以解决某个已被提出的问题
 - 新颖性在于合理地选择各个技术部件，达到最终的目标
 - 这类工作往往难以发表在质量较高的期刊或会议上

给出解决方案的过程

- 提出观点
- 给出验证

2 提出观点

提出观点

- 在确切地知道答案之前，需要一个暂时答案（working answer）
 - 即使你认为这个暂时性答案将被其他答案所取代
 - 通过这个答案来帮你找出相关的数据
 - 通过这个观点来帮你整理出用来支持观点的论证
- 试着拟出目前所能了解的范围内最好、最完整的研究观点

如果你过于相信，你就永远无法注意到缺点。

如果你怀疑太多，你将不会开始。



理查德·卫斯里·汉明
Richard Hamming

美国数学家，
图灵奖得主。

克劳德·艾尔伍德·香农
Claude Elwood Shannon



美国数学家、电子工程师和密码学家，被誉为信息论的创始人。

勇气是香农拥有的最重要的事物之一

——Richard Wesley Hamming

观点

- 通过三个问题检验观点：
 - 要提出哪一类观点？
 - 是否能具体地陈述？
 - 是否是重要的观点？

观点的类型

- 描述问题的原因，要求受众相信什么

“2001—2002年经济衰退的部分原因是信息系统投资过多，这些投资没能像预期的那样提高生产力。”

- 给出行动方案，告诉受众可以去做什么

“调整投资结构，降低对信息系统的投资，注重实业发展。”

观点的明确性

- 观点越明确，受众越能判定它具有实质性的内容
- 观点越明确，越容易规划一个支持它的实质论证



观点的明确性

- 承认限定条件
 - 没有任何一个观点是没有限定条件的
- 适当的确定性
 - 谨慎的作者用模糊的修饰词来承认研究的局限
 - 完全确定的字眼如all、none、every、always、never.....不会留下好印象
 - 但太多的模糊限制语，看起来会有胆怯或不够确定的感觉

观点的明确性

- 具体明确的语言

- ✗ 电视媒体夸大犯罪率的估计

- ✓ 地方电视媒体有关暴力画面的报道，导致民众将他们小区的犯罪率与他们和家人可能遭到危险的概率至少高估**150%**

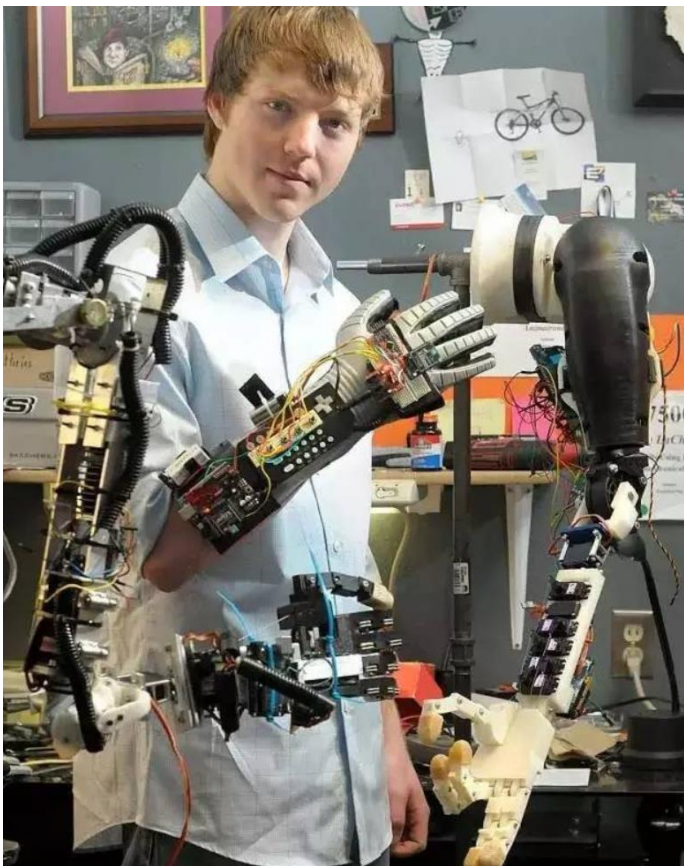
观点的明确性

- 具体明确的逻辑
 - 以虽然（although）或即使（even though）开头的从句来引出观点
 - 以因为（because）开头的“原因从句”（reason-clause）来结束观点

观点的意义

- 从先前已被探究的题目中得到的新信息
- 解决过去令人困扰的问题的研究（更好）
- 颠覆了早已被认为是毋庸置疑的问题的新知识（最好）

3 给出验证



Easton LaChappelle
科技鬼才、天才创客

喜欢变形金刚，为了帮助
没有手臂的小女孩研发了
机器手

停止空想开始行动！

论证结果

- 组织数据内容
 - 汇集论证的要素
 - 以符合受众知识和需求的方式分类整理
- 论证模式
 - 不是赢家与输家之间的争执，而是深思熟虑的同事间友善的对谈
 - 合作探索具有争议性的话题，并且双方同意解决这个议题是重要的

研究论证的要素

- 什么是你的观点(**claim**)?
- 什么理由能支持你的观点(**reasons**)?
- 什么证据(**evidence**)能支持你的理由?
- 你承认(**acknowledge**)这些不同的替代方案/复杂的情形/或反驳么? 你如何做出回应(**respond**)呢?
- 什么原理/论据(**principle/warrant**)能证明你的理由和观点之间的关联呢?

将观点建立在理由的基础上

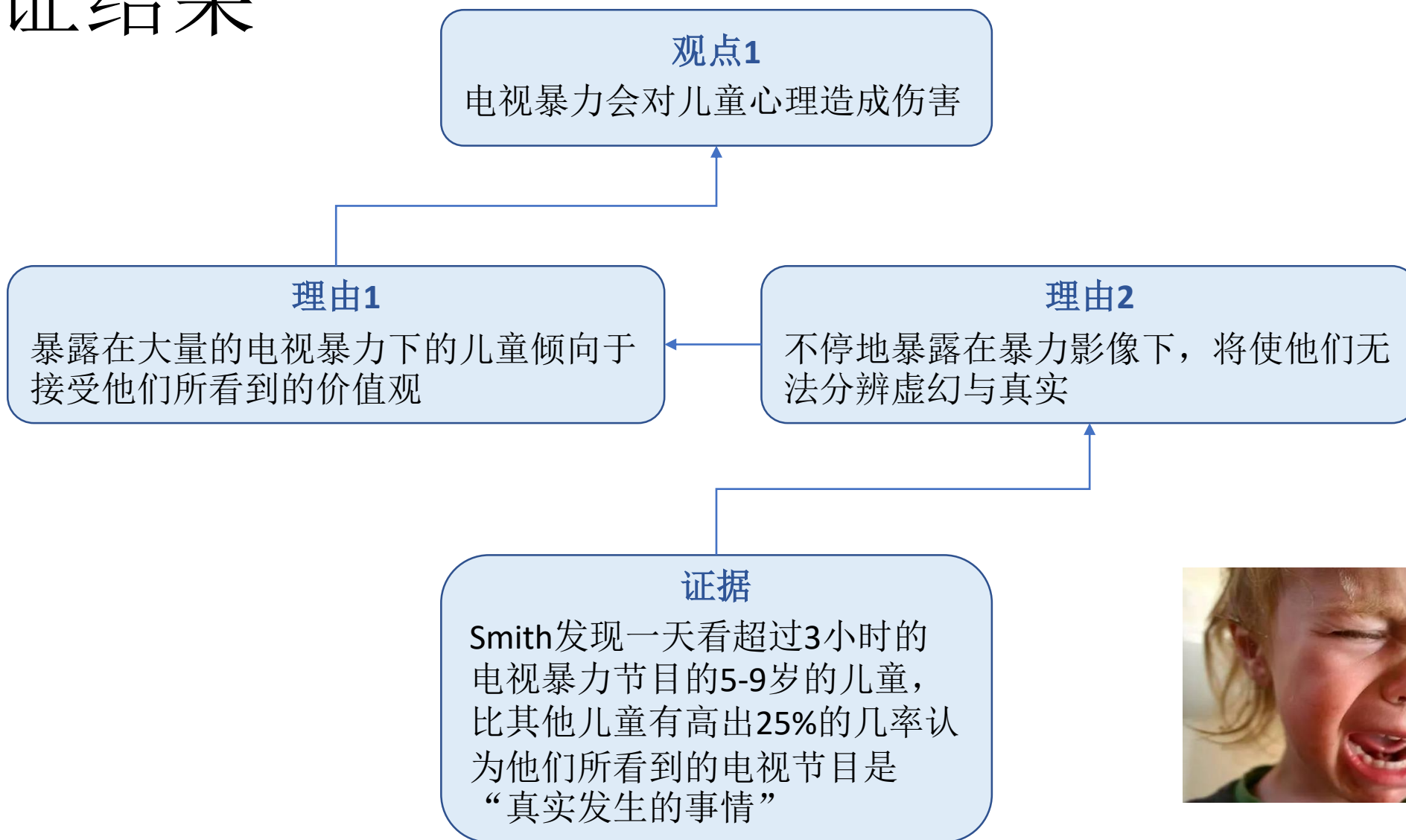
- 观点是研究报告的核心，也是研究问题的答案
- 理由用一两句话解释为什么应该接受观点
- 一个句子可能同时是理由又是观点

将理由建立在证据的基础上

- 平常的交谈常用理由去支持观点：
 - 我们应该走了，（观点）
 - 因为似乎要下雨了。（理由）
- 而不会去问要下雨的证据是什么
 - 气象学家会说：那片云云底阴暗，是积雨云，会产生降雨。（证据）



论证结果



为理由的关联赋予论据

- 即使受众同意证据足够充分，他们可能仍不接受你的观点
- 被质疑的可能不是理由的真实性，而是理由与观点之间的关联性
- 为了回答问题，必须提出一般性的原则，以显示为什么相信这个特定理由与这个特定观点有关联

论证结果

- 核心架构

支持 **观点** 的 **理由** 建立于 **证据** 的基础之上

论证结果

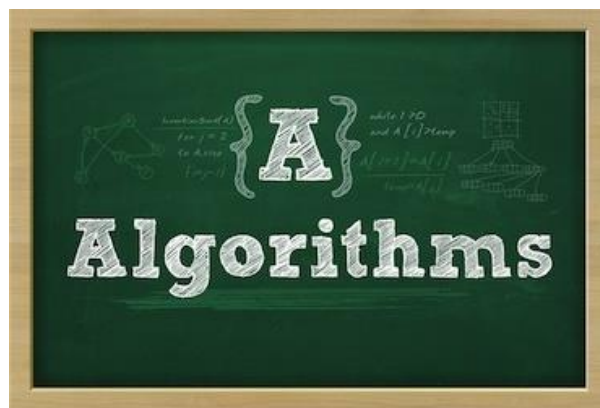
- 回到之前的“一句话描述研究目的”
 - “在搜索算法中加入社交网络这一因素会比不考虑社交网络的搜索算法带来更加精准的搜索结果”
- 假设现在你正在进行这项工作，如何证明他的意义和优越性？

论证结果

- 理论证明
- 实验证明
- 对结果进行人为判断
- 绘制图表和分析

理论证明

- 例如：通过理论证明，无论PageRank产生怎样的排序结果，你的算法产生的结果总不会差过他
- 需要提供广泛认可的评价标准
 - 排序算法的评价标准：AUC、NDCG等
 - 结果与你的理论证明保持一致
 - 也就是要让你自己的比较好
- 对于偏理论型的研究，每一步都要有相应定理和推论的支持



实验证明

- 通过实验与其他工作的结果进行比较
- 寻找合适的实验数据
- 参数的设计和调节



寻找合适的数据

- 可以寻找对应问题和算法常用的公开数据集
 - 例如用户搜索历史数据
 - 通常很大，且包含噪声
- 清理原始数据
 - 确保抽取的数据符合元数据分布
 - 均匀采样
 - 例如每个月均匀分布的采样
 - 对于较为复杂的数据，可以使用蒙特卡罗（Monte Carlo）方法
 - 根据输入参数反复运行模拟从而确保抽取的数据符合元数据分布



参数的设计和调节

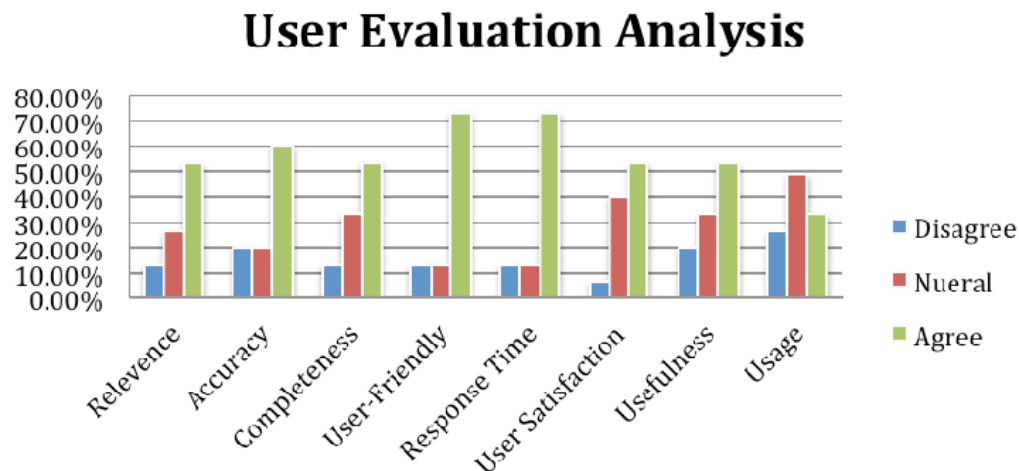
- 自由参数/独立参数
 - 不依赖其他任何参数
 - 对实验结果的影响能直接观察到
 - 含有多个自由参数时要控制变量
 - 绘制以单一参数为横轴，结果评价为纵轴的曲线

对结果进行人为判断

- 除了恰当的评价标准、与其他工作的比较之外，有时还需要对结果进行人为的判断和评价
- 被试的选取
 - 确保具有公正性、独立性以及不同背景
 - 找些心理学方面的书籍作为参考

绘制图表和分析

- 将实验结果绘制成简洁的图表



- 对图表进行文字性的分析
 - 让阅读者从统计学角度指导实验结果有无显著性差异
 - 将论文中的实验结果从专业知识的角度进行解释
 - 说明各组实验的意义

统计学角度

- 显著性测试（significance test）/假设检验
 - 先对总体特征做出假设
 - 根据统计推理对假设做出接受/拒绝的判断
 - T检验
 - 用于样本含量小、总体标准差未知的正态分布
- p值检验
 - 以期望值p作为判断依据

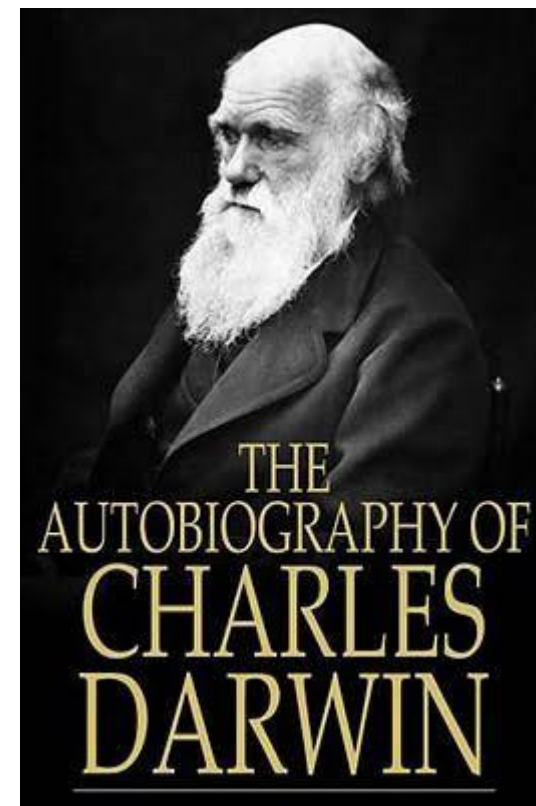
回应质疑

- 研究者以有证据基础的理由来支持观点
- 但是，受众可能会
 - 想到其他证据
 - 对给出的证据有不同解读
 - 从相同的证据中引出不同的结论

很有必要记录下所有可能与自己信条相矛盾的每一条证据。如果不这么做，它们很快就会从头脑中消失。

当你发现明显的缺陷时，你必须敏感地意识到并持续关注这些事情，注意如何解释它们，或者如何调整理论以适应它们。这些往往是伟大的贡献。

——达尔文



防守式论证

- 预估受众提出的问题
- 承认与回应其中最重要的问题
- 具体做法
 - 阅读相关文献，留意论文里提及的问题及其回应方式
 - 和旁人谈论观点，请他们指出问题或者是否有其他的替代方案



承认与回应

证据

Smith发现一天看超过3小时的电视暴力节目的5-9岁的儿童，比其他儿童有高出25%的几率认为他们所看到的电视节目是“真实发生的事情”

承认

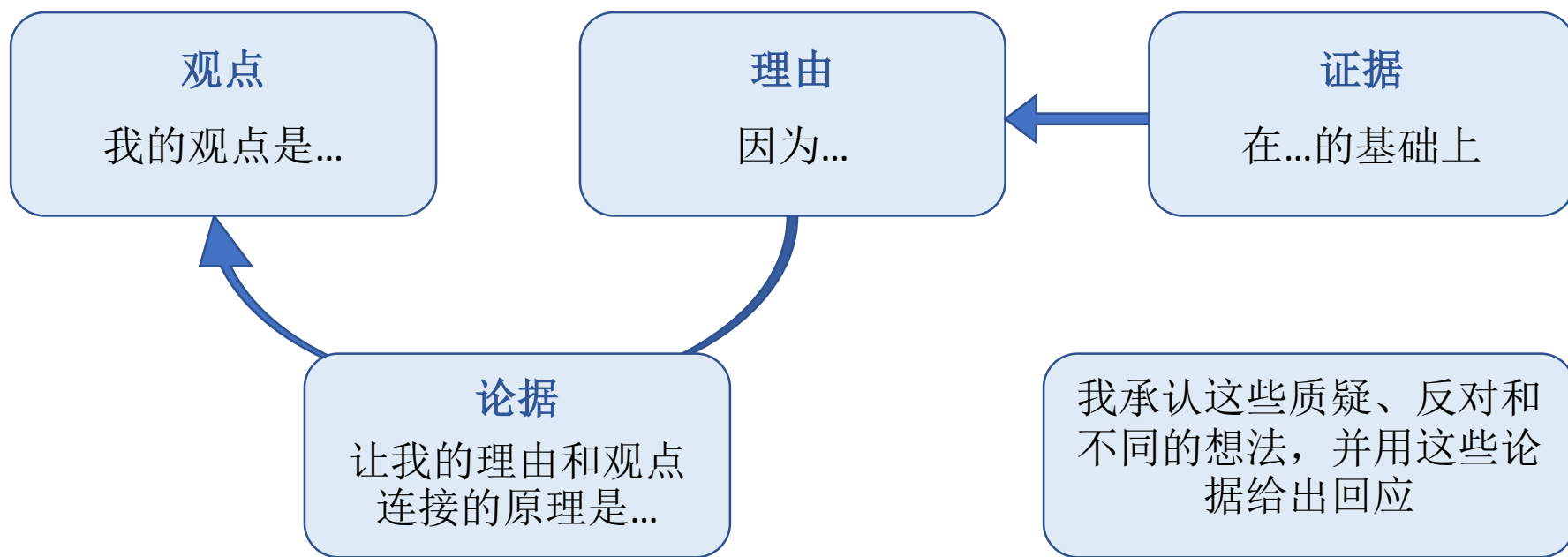
当然，倾向于观看大量的暴力性的节目的儿童可能已具暴力的价值观



回应

但是Jones发现不论儿童有无暴力倾向，他们都易被具有暴力性的电视节目吸引。

研究论证的组成



4 一些误区

解决问题的误区

- 其一：将科研等同于解题，找到一个可行方案就停止搜索
 - 再往前走一步可能会有更好的发现或结果

用炭棒做灯丝，发光刺眼、耗电大，使用寿命也很短。爱迪生决定要发明一种灯光柔和的电灯。爱迪生看到老朋友身上的棉线外套，突然大叫：“棉线，为什么不用棉线呢！”

用棉线做灯丝的灯泡并不刺眼，爱迪生取得了一个重大突破，新的问题又来了：怎样才能让这个灯亮得更久呢？爱迪生看到平时用的竹扇面，拿竹丝来做灯丝一试，发现了更耐用的灯丝。



具体策略

- 流利思考
 - 有了想法，多想一点，想深一点
 - 保持思路不被打断，创造更多可能性和解决方案

*I think the deeper you go into questions, the deeper or more
interesting the questions get. And I think that's the job of art.*

——Andre Dubus



Andre Dubus
短篇作家，小说家

解决问题的误区

- 其二：过于担心方法的缺点而踟蹰不前
 - 对科研文章，新颖度的评价远大于是否实用和适用

[PDF] Visualizing data using t-SNE.

[L Van der Maaten](#), [G Hinton](#) - Journal of machine learning research, 2008 - jmlr.org

We present a new technique called “t-SNE” that visualizes high-dimensional data by giving each datapoint a location in a two or three-dimensional map. The technique is a variation of Stochastic Neighbor Embedding (Hinton and Roweis, 2002) that is much easier to optimize ...

☆ 被引用次数: 22339 相关文章 所有 57 个版本

[HTML] How to use t-SNE effectively

[M Wattenberg](#), [F Viégas](#), [I Johnson](#) - Distill, 2016 - distill.pub

A popular method for exploring high-dimensional data is something called t-SNE, introduced by van der Maaten and Hinton in 2008 [1]. The technique has become widespread in the field of machine learning, since it has an almost magical ability to create compelling two-dimensional ...

☆ 被引用次数: 422 相关文章 所有 3 个版本

[PDF] Accelerating t-SNE using tree-based algorithms

[L Van Der Maaten](#) - The Journal of Machine Learning Research, 2014 - jmlr.org

The paper investigates the acceleration of t-SNE—an embedding technique that is commonly used for the visualization of high-dimensional data in scatter plots—using two treebased algorithms. In particular, the paper develops variants of the Barnes-Hut algorithm ...

☆ 被引用次数: 1715 相关文章 所有 19 个版本

后续工作不断改进原有工作的不足

在行动上尽可能坚定果断，一旦选定某种看法，哪怕它十分可疑，也毫不动摇地坚决遵循，就像它十分可靠一样。

有一条非常可靠的真理，就是在无法分明分辨哪种看法最正确的时候必须遵从或然性最大的看法。

勒内·笛卡尔
René Descartes



法国著名哲学家、
数学家、物理学家。
解析几何之父。

解决问题的误区

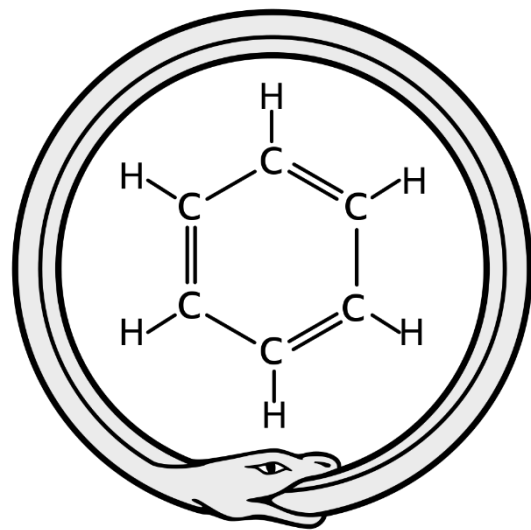
- 其三：自认为解决方法很完美
 - 以谦虚的态度做研究
 - “To the best of our knowledge”
 - “As far as we know”
 - “It seems that the previous work”
 - 不夸大优点，不忽视缺点

解决问题的误区

- 其四：找不到解决方案时更加专注于研究问题本身
 - 需要开拓视野，换个角度看问题



跳出圈子，从其他事物找灵感



蛇咬尾→苯的环状结构



撕碎的报纸→大陆漂移说

具体策略

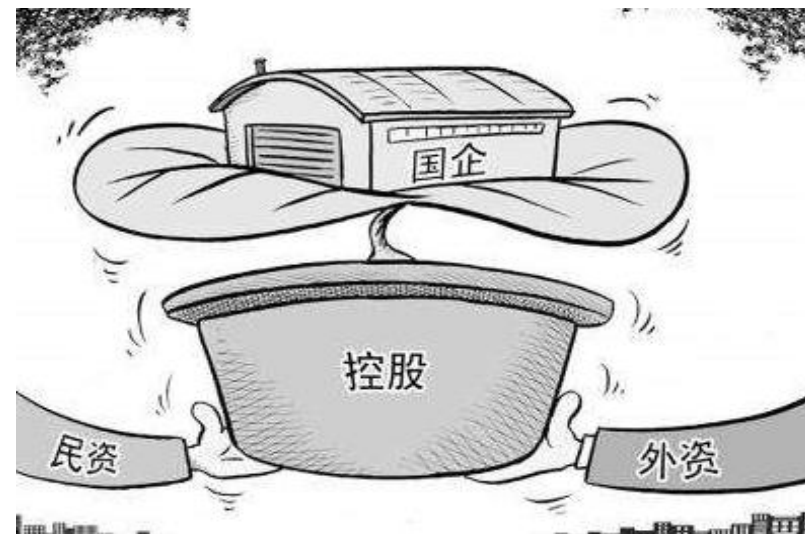
- 了解其他领域的研究是否存在相似问题？
- 他们是如何解决的？
- 能否迁移？

具体策略

- 尝试新组合
 - 进行适应性优化
 - 检验成果确实有效



荔枝蘸酱油



社会主义市场经济体制