数据挖掘与信息检索实验室

Data Mining & Information Retrieval Laboratory

科研论文的撰写经验分享

——座谈提纲@肇庆学院

蔡瑞初

cairuichu@gmail.com

Outline

- ☑ 为什么要写论文?
- ☑ 如何选题?
- ☑ 如何研究?
- ☑ 如何写作?

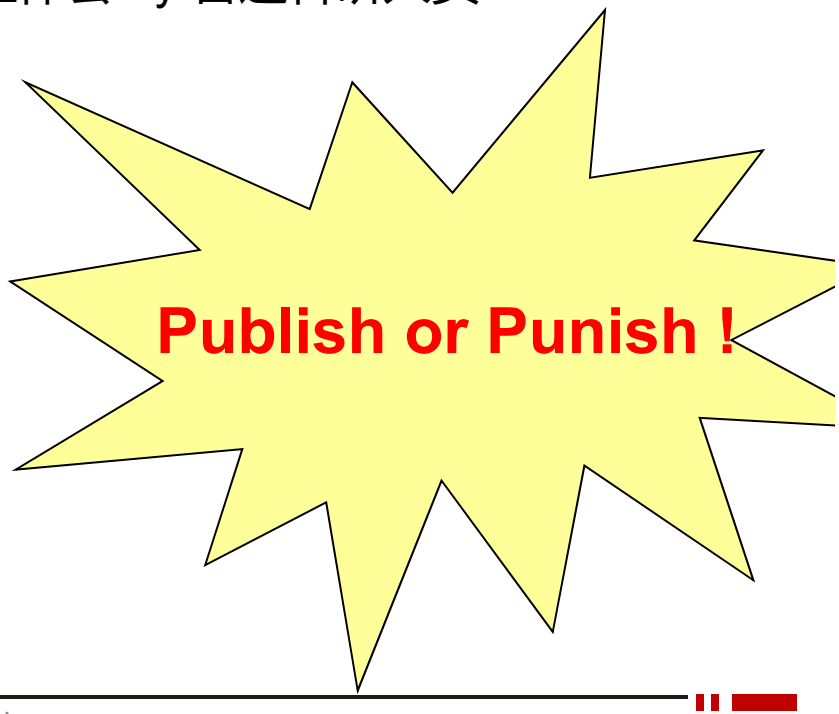
为什么要写论文?

为什么要写论文？——三个层次

☑ 职责要求：评职称、拿项目、学生要毕业 by 普通科研人员

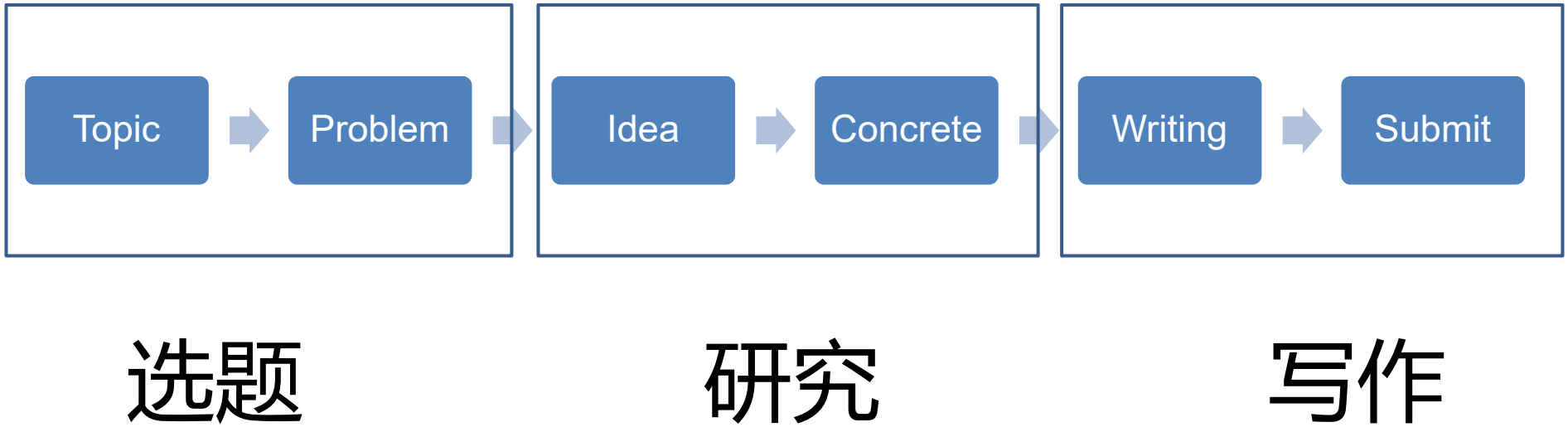
☑ 学术交流：与同行交流、分享研究成果或经验体会 by 普通科研人员

☑ 彻底解决问题：阻止别人发论文 by 某院士



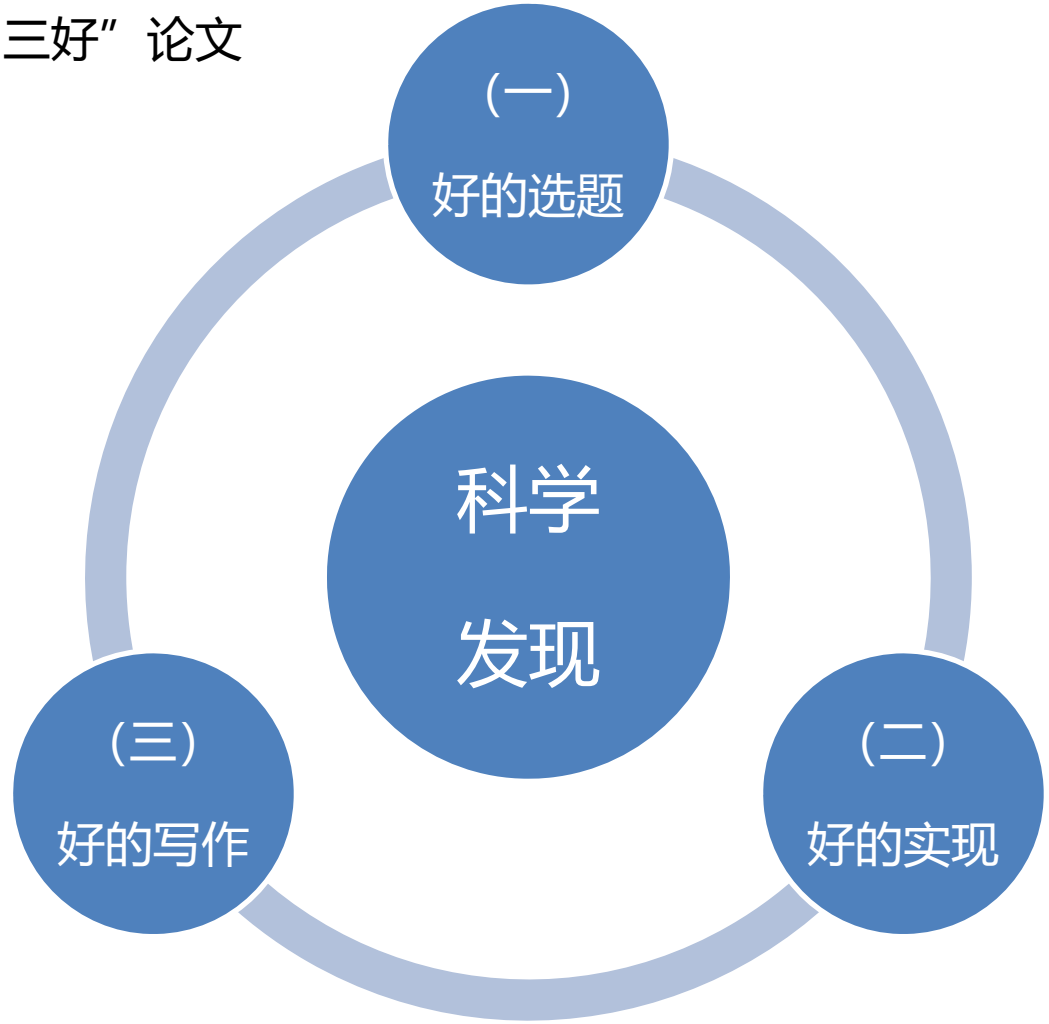
为什么要写论文？——从研究流程来看

☑ 论文是研究的自然结果



为什么要写论文？——好论文长什么样？

- ☑ 符合论文的初心：分享科学发现
- ☑ 一流的工作，“三好”论文

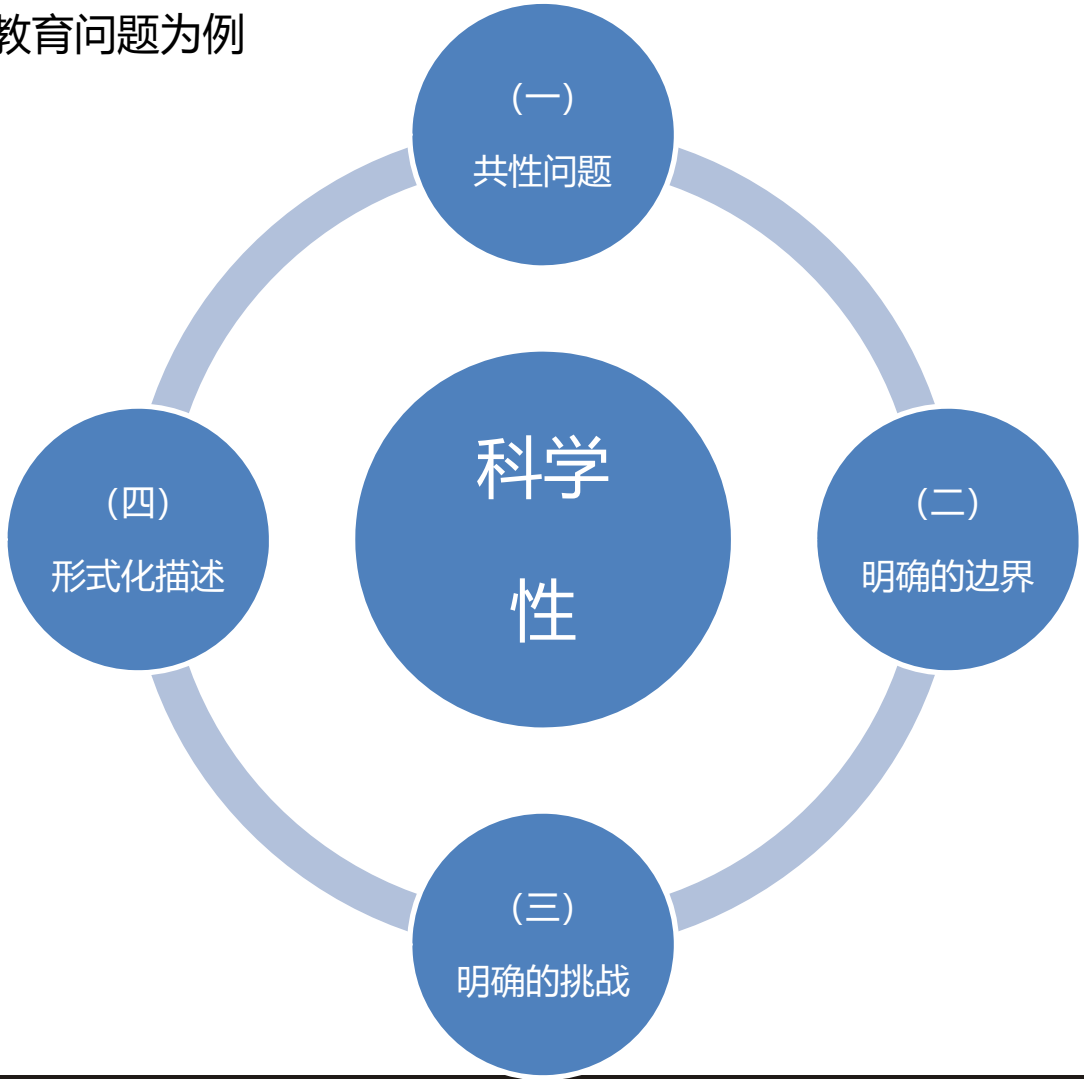


如何选题?

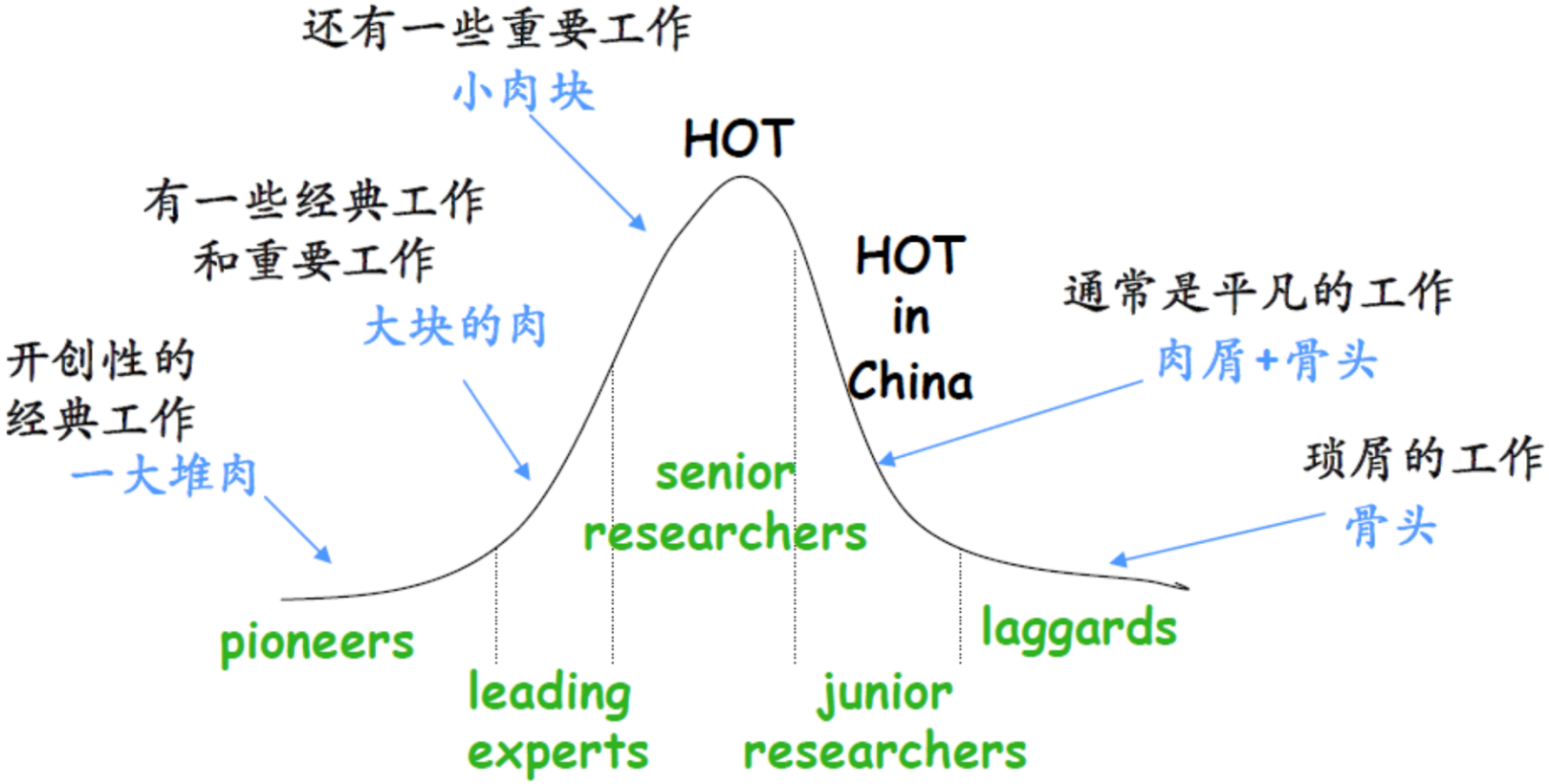
如何选题？ ——什么是选题？

☑ 如何定义选题？科学性

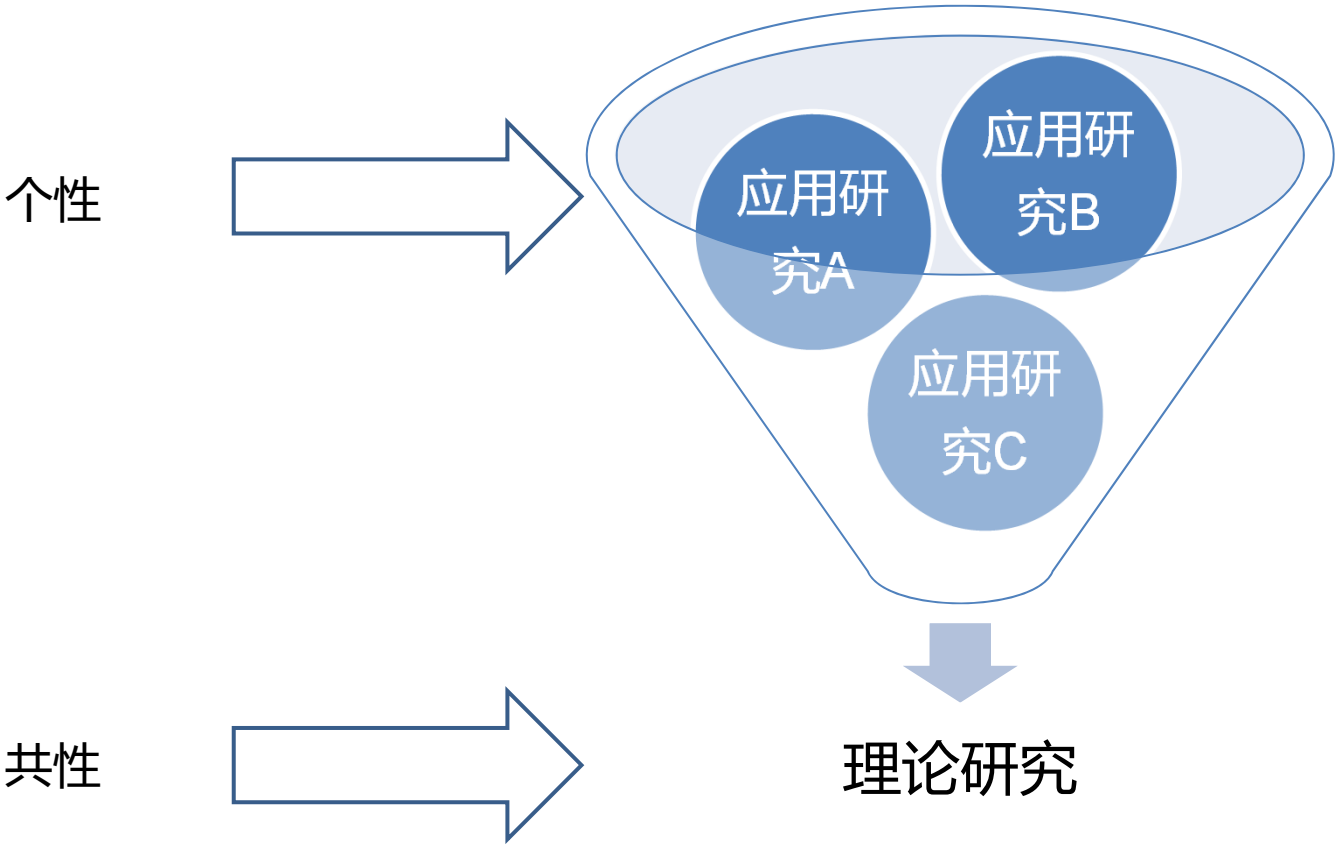
- 家里的小魔王的教育问题为例



如何选题？ ——前沿 V.S. 经典


还有一些重要工作
小肉块
HOT
有一些经典工作
和重要工作
大块的肉
HOT
in
China
通常是平凡的工作
肉屑+骨头
开创性的
经典工作
一大堆肉
senior
researchers
琐屑的工作
骨头
pioneers
leading
experts
junior
researchers
laggards

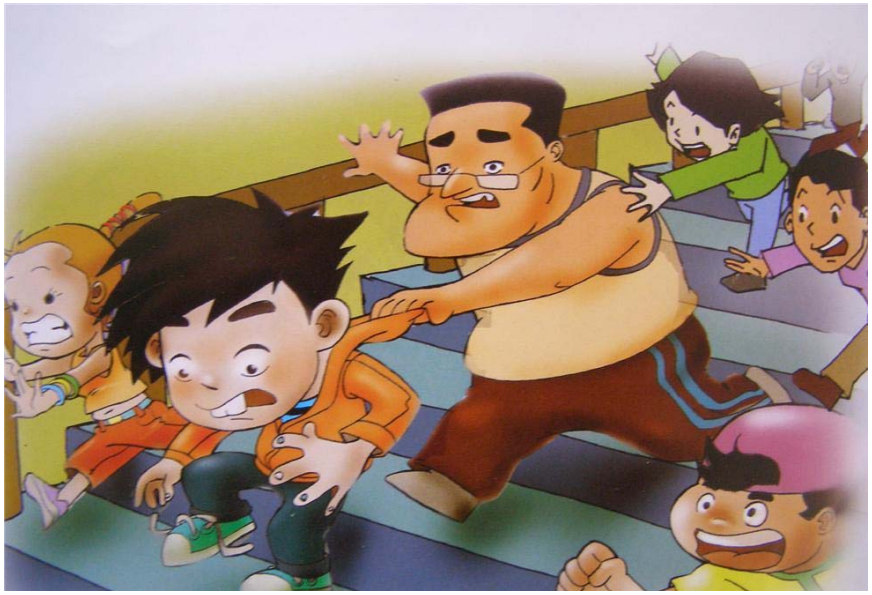
如何选题？ ——理论V.S.应用



如何选题？ ——中流砥柱 V.S. 随波逐流

☑ 对你的游泳速度有信心吗？

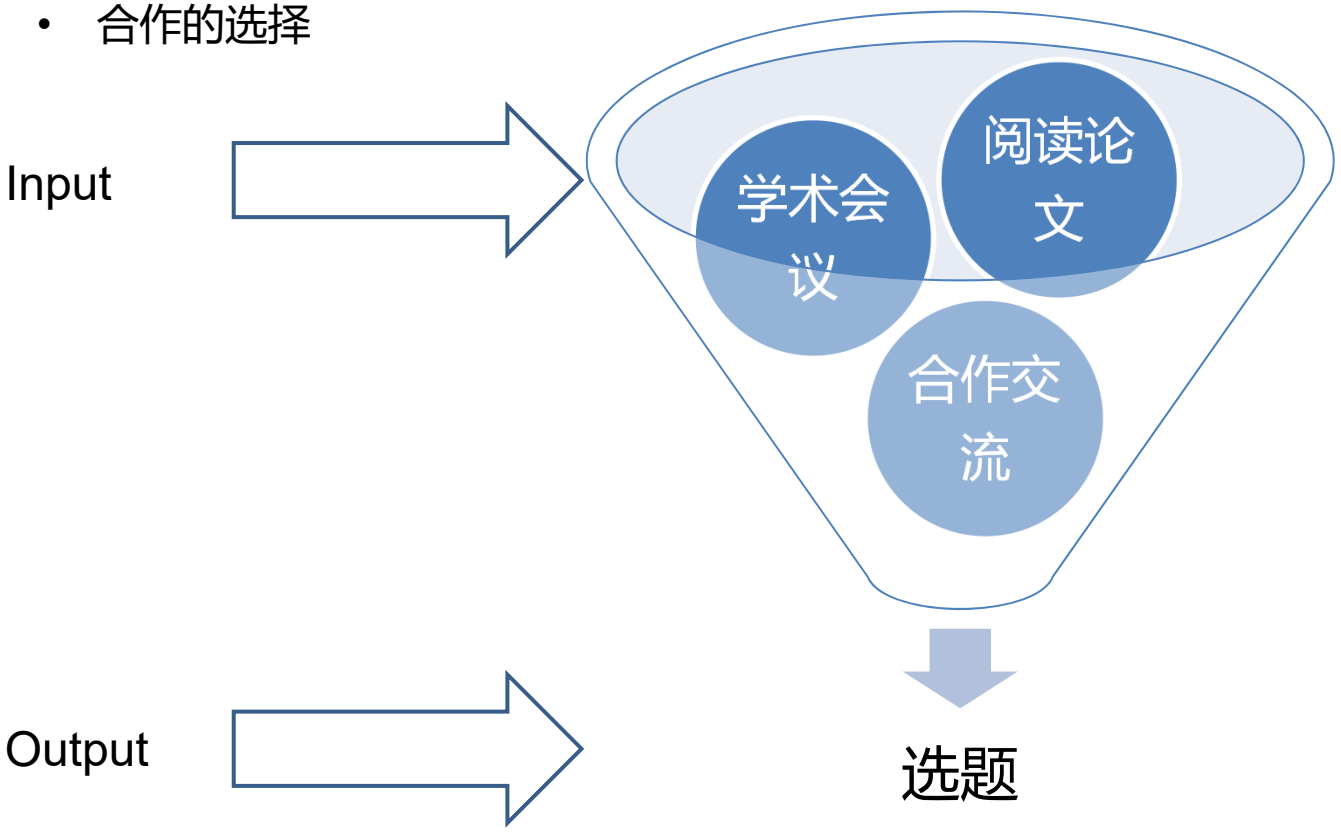
- 踩踏风险、理性看待



如何选题？ ——没的选怎么办？

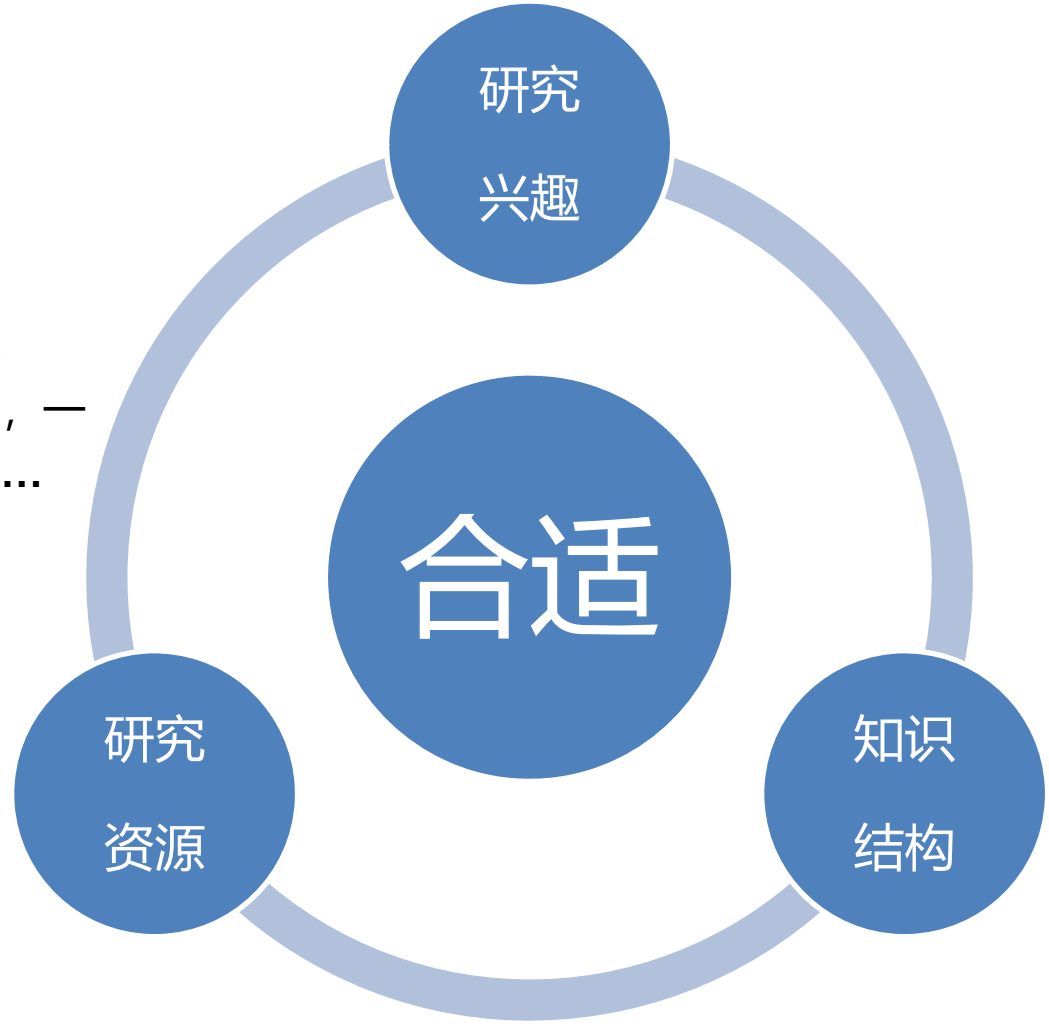
☑ Input的选择

- 吃的是草挤的是奶？牛人！
- 普通人？Annuals of statistics → JMLR → A会→ IEEE Trans → SCI → 核心期刊
- 会议的选择
- 合作的选择



如何选题？ ——从个人角度

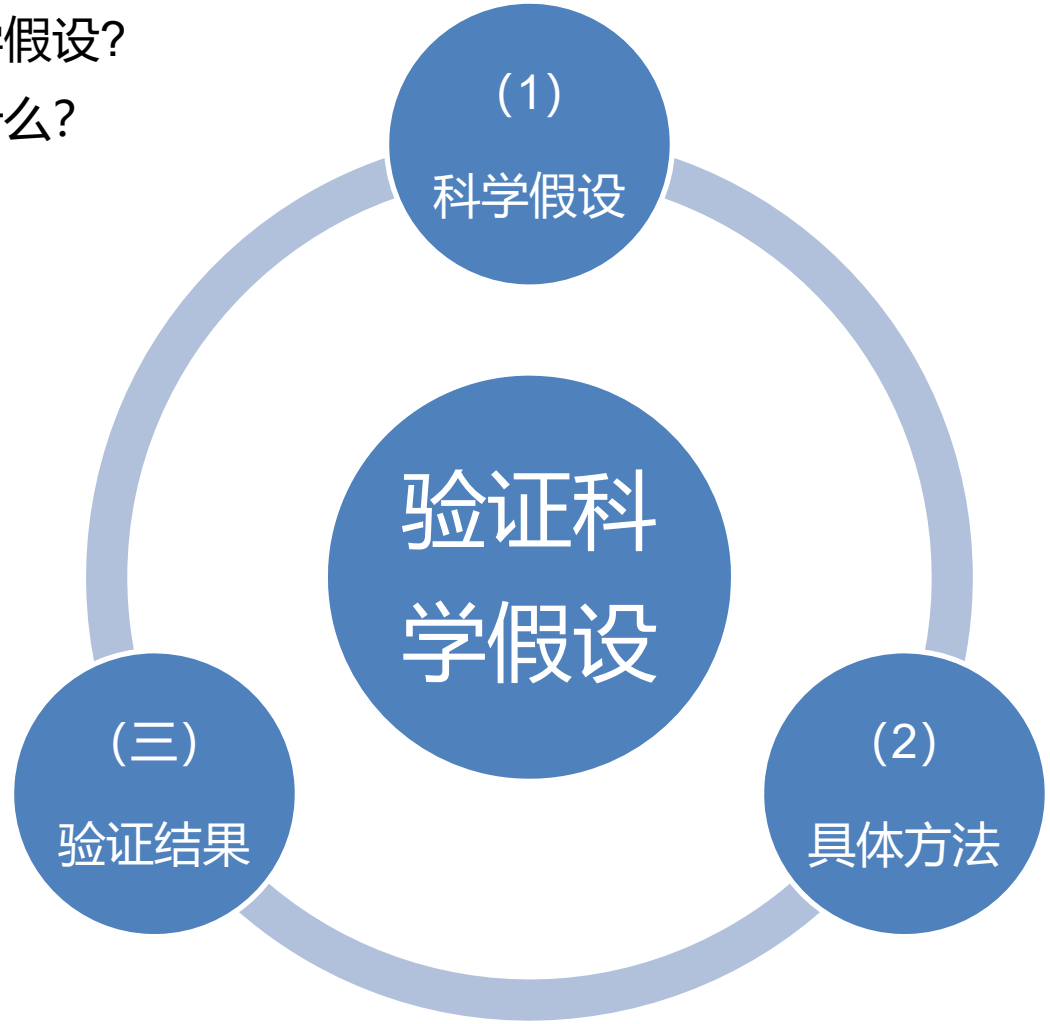
- ☑ 选择最适合你的**topic:**
- ☑ 研究兴趣
  - 这是最重要的!
  - 国外的CV
- ☑ 知识结构
  - 基础，但是可以克服
  - 没有必要的知识积累，一切从头开始的话，……
- ☑ 研究资源
  - 重要，但是可以获得
  - 例如数据



如何研究?

如何研究？——三个阶段

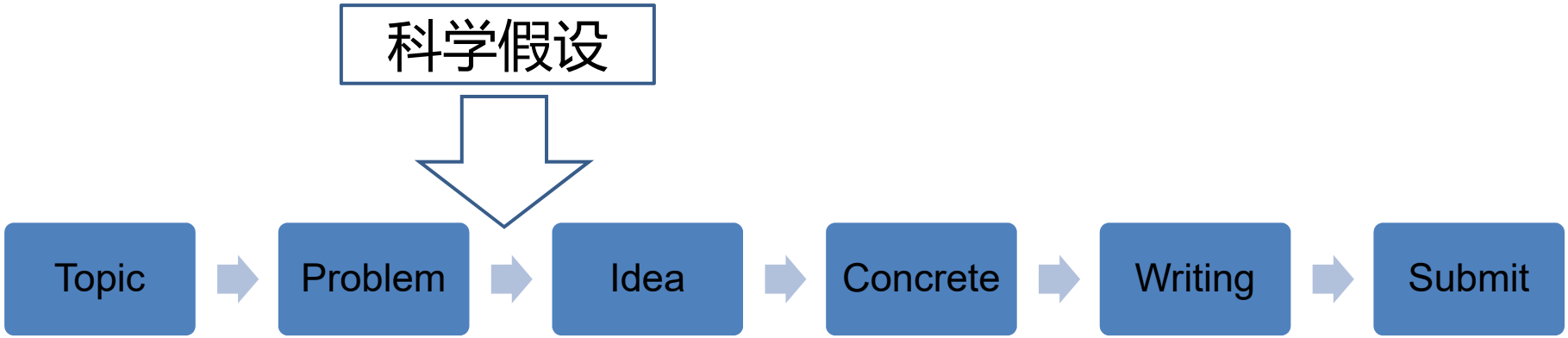
- ☑ 科学假设是什么？
- ☑ 怎么去验证科学假设？
- ☑ 验证的结果是什么？



如何研究？ ——科学假设

☑ 为什么要有科学假设？

- 承接problem & idea



☑ 评判标准：

- Rubbish in Rubbish out！
- 要创新、要独到

☑ 来源问题：对于问题的认识、观察、

如何研究？ ——具体方法

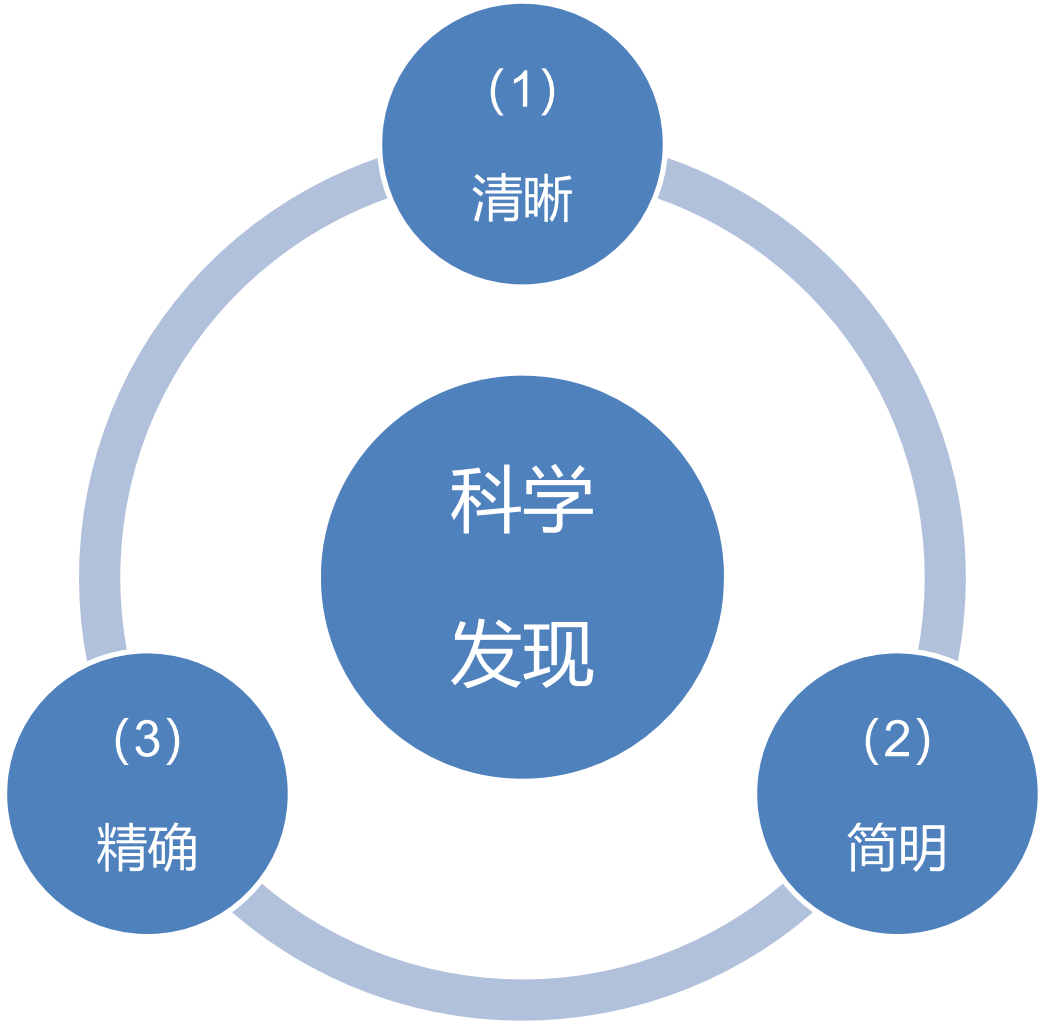
- ☑ 评判标准：
  - 要创新、要solid
  - 一流：新的方法（新的定义是什么？）
  - 二流：旧的方法改进版
  - 三流：旧的方法
- ☑ 力求：
  - 新的、合理、逻辑严密的解决方法
  - 理论支持的方法
- ☑ 避免：
  - 有逻辑漏洞
- ☑ 讨论
  - 用了旧的方法，论文的方法部分怎么写？

如何研究？ ——验证结果

- ☑ 评判标准：
  - 要验证科学假设、要比对比方法好、要可验证
- ☑ 力求：
  - 实验方案周全、仔细
  - 结论客观
  - 深刻的分析：为什么好？为什么不好？好与不好能否验证假设？
  - 结果可重复（其他学者也能使用的数据、也有同样环境）
- ☑ 避免：
  - 只有结果，没有分析，没有结论
  - 只追求结果，没有验证科学假设
- ☑ 讨论/误区：
  - 效果不好的论文能发吗？（效果不好只有一种标准，不能验证科学假设）

如何写作?

如何写作?



如何写作？——想想你的目标读者/目标杂志

☑ 确定投稿目标：

- 根据工作的水准，挑选合适的发表源
- 低投—遗憾，高投—延误发表

☑ 针对不同的发表源可能有不同的写法

- 读几篇对应期刊的论文，看他的偏好

☑ 计算机领域

- 会议有其重要特殊地位， CCF A类列表

如何写作?

☑ 题目

- 用最少的字、最准确地概括论文内容。
- 力求：简短、明确、完整、有吸引力
- 避免：冗长、模糊、拔高、空洞

☑ 简朴style： 创新+手段+问题

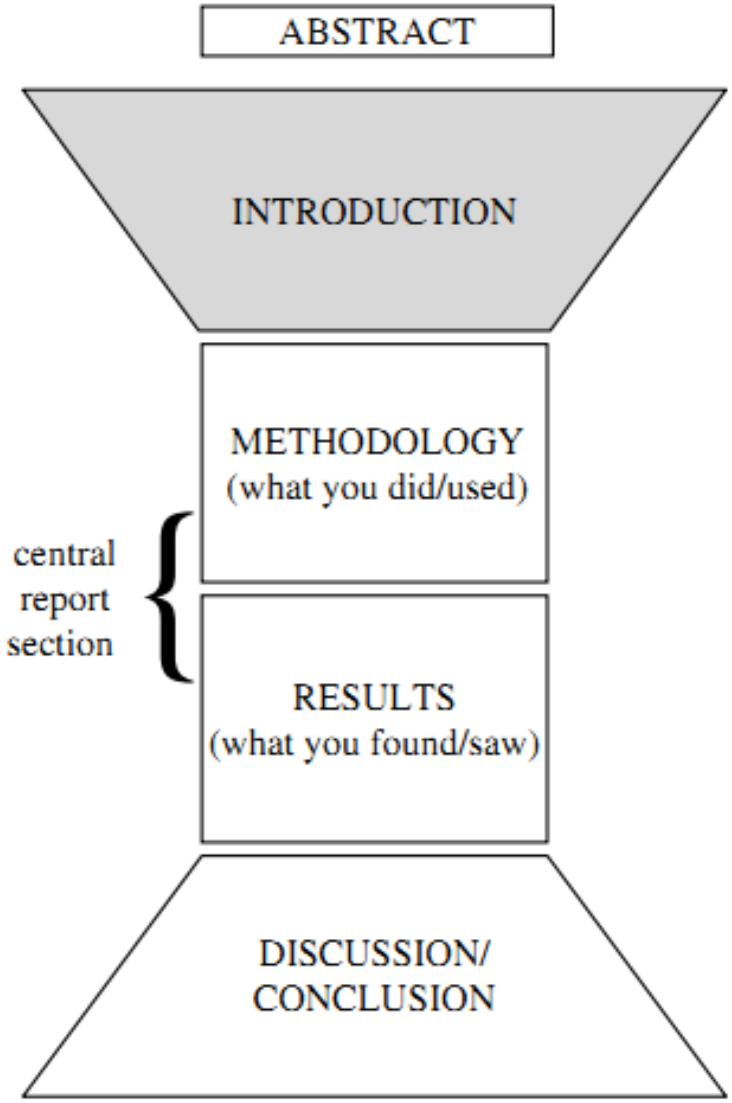
Learning disentangled semantic representation for domain adaptation

☑ 吸引眼球

What is unequal among the equals? Ranking equivalent rules from gene expression data

如何写作？ —— 架构

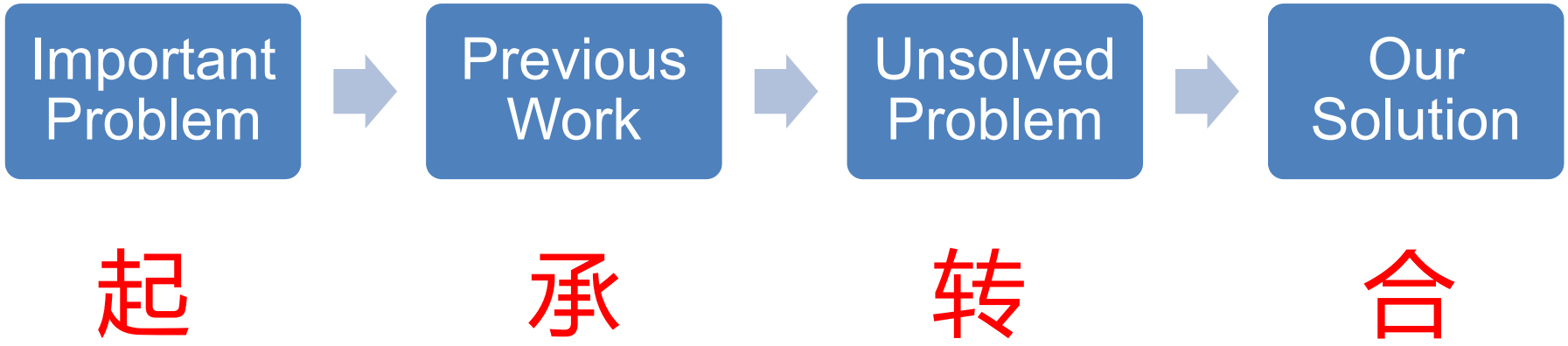
☑ 基本框架


ABSTRACT
INTRODUCTION
METHODOLOGY
(what you did/used)
central
report
section
RESULTS
(what you found/saw)
DISCUSSION/
CONCLUSION

如何写作？ —— Introduction

☑ 目的：论证科学问题的重要性和解决方案的合理性

☑ Style



☑ 力求： 逻辑清晰、论证到位

☑ 避免：眉毛胡子一把抓、攻击对手不到位

如何写作？ —— Introduction

起：Domain adaptation is an important but challenging task……

承：An essential approach in unsupervised domain adaptation is to understand what the domain-invariant representation across the domains is and how to find it [Zhang et al., 2013; Pan et al., 2011; …..

转：Due to the complex manifold structures underlying the data distributions, these methods mainly suffer a false alignment problem [Pei et al., 2018; Xie et al., 2018]…

合：Motivated by the example of Figure 1(b), the underlying cause of the false alignment problem is the entanglement of the semantic and domain information…

Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)

Learning Disentangled Semantic Representation for Domain Adaptation

Ruichu Cai[1], Zijian Li[1], Pengfei Wei[2], Jie Qiao[1], Kun Zhang[3] and Zhifeng Hao[4]
[1]School of Computers, Guangdong University of Technology, China
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Department of Philosophy, Carnegie Mellon University, USA
[4]School of Mathematics and Big Data, Foshan University, China
cairuichu@gdut.edu.cn, leizigin@gmail.com, wpf89928@gmail.com, kunz1@cmu.edu,
qiaojie.chn@gmail.com, zfhao@gdut.edu.cn

Abstract

Domain adaptation is an important but challenging task. Most of the existing domain adaptation methods struggle to extract the domain-invariant representation on the feature space with entangling domain information and semantic information. Different from previous efforts on the entangled feature space, we aim to extract the domain invariant semantic information in the latent disentangled semantic representation (DSR) of the data. In DSR, we assume the data generation process is controlled by two independent sets of variables, i.e., the semantic latent variables and the domain latent variables. Under the above assumption, we employ a variational auto-encoder to reconstruct the semantic latent variables and domain latent variables behind the data. We further devise a dual adversarial network to disentangle these two sets of reconstructed latent variables. The disentangled semantic latent variables are finally adapted across the domains. Experimental studies testify that our model yields state-of-the-art performance on several domain adaptation benchmark datasets.

1 Introduction

Domain adaptation is an important but challenging task. Since the acquisition of a large labeled data is usually either expensive or impractical, how to train on the unlabeled target domain with the help of labeled source domain has become a particular focus. However, this learning scheme suffers from a well-known phenomenon named *domain shift*, leading an urging motivation in building an adaptive classifier that can efficiently transfer the source labeled data under the domain shift, this problem is also known as *unsupervised domain adaptation*.

An essential approach in unsupervised domain adaptation is to understand what the domain-invariant representation across the domains is and how to find it [Zhang *et al.*, 2013; Pan *et al.*, 2011; Gong *et al.*, 2012]. Typical methodologies explored in the literature include the feature alignment approaches that extract domain-invariant representation by minimizing the discrepancy between the feature distributions inside deep feed-forward architectures [Tzeng *et al.*, 2014; Long *et al.*, 2015; 2016; 2017], and the adversarial learning approaches that extract the representation by deceiving the domain discriminators [Tzeng *et al.*, 2015; Ganin and Lempitsky, 2015; Ganin *et al.*, 2016; Long *et al.*, 2018]. Recently, a fine-grained semantic alignment has also been proposed in order to extract domain-invariant representation under the consideration of the semantic information [Xie *et al.*, 2018; Zhang *et al.*, 2018; Chen *et al.*, 2018]. However, most of them require target pseudo labels in order to minimize the discrepancies across domains within the same labels, and thus resulting the error accumulation due to the uncertainty of the pseudo-labeling accuracy.

Due to the complex manifold structures underlying the data distributions, these methods mainly suffer a false alignment problem [Pei *et al.*, 2018; Xie *et al.*, 2018]. As shown in Figure 1(a), the data generation process is controlled by the disentangled domain latent variables and semantic latent variables in the latent manifold. The ideal *semantic* position of the two types of labels – pig and hair drier – should be placed relatively upper and lower on the manifold according to the semantic axis, and the ideal *domain* position within the same labels should be placed in the left and right according to the domain axis. However, as shown in Figure 1(b), once the domain information is not completely removed, samples are distorted on the feature manifold, leading to the false alignment problem, e.g., the Peppa Pig looks like a pink hair-drier and thus the feature of the Peppa Pig might be near to that of the hair-drier in the distorted feature manifold.

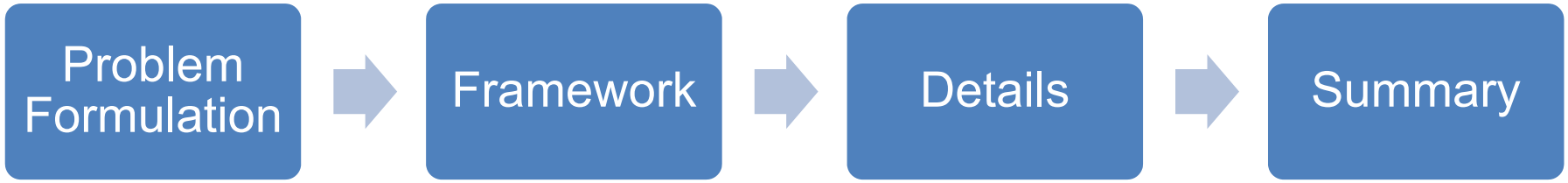
Motivated by the example of Figure 1(b), the underlying cause of the false alignment problem is the entanglement of the semantic and domain information. More specifically, samples are controlled by two sets of independent latent variables $\mathbf{z}_y$ and $\mathbf{z}_d$. However, these two sets of latent variables are highly tangled and distorted on the high dimensional feature manifold space. It is very challenging to remove the domain information while preserving the semantic information on a complex tangled feature manifold space.

In this work, motivated by the disentanglement property of the multiple explanatory factors in the representation learning literature [Bengio *et al.*, 2013; Dinh *et al.*, 2014], we propose a Disentangle Semantic Representation learning model (DSR in short) by assuming the independence between the semantic variables z_y and the domain variables z_d. Our DSR reconstructs the disentangled latent space and simultaneously

如何写作？ —— Methodology

☑ 目的：介绍解决方案，论证方案的科学合理性

☑ Style



起 总 分 总

☑ 力求： 方案合理、逻辑清晰、

☑ 避免：知其然不知其所以然

如何写作？—— Methodology

起：

In this work, we focus on the unsupervised domain adaptation problem that uses the labeled samples $D_S = \{x_i^S, y_i^S\}_{i=1}^{n_S}$ on the source domain to classify the unlabeled samples $D_T = \{x_j^T\}_{j=1}^{n_T}$ on the target domain. The goal of this paper is to understand:(1) what the domain-invariant representation across domains is, and (2) how to design a framework that can extract such a domain-invariant representation.

总：

Regarding the second point, with the above data generative mechanism, we propose a disentangled semantic representation (DSR) domain adaptation framework by first reconstructing the two independent latent variables via variational auto-encoder and then disentangling them through a dual adversarial training network. The key structure of the proposed framework is given in Figure. 3.

分：

3.1 Semantic Latent Variables Reconstruction

3.2 Semantic Latent Variables Disentanglement

总：

3.3 Model Summary

By combining the reconstruction and the disentanglement, we summarize the model as follows.

如何写作？ —— Results/Experiments

☑ 目的：介绍实验方案与验证结果，验证方案的科学合理性

☑ Style

☑ 力求： 细节充分、实验可以重复、分析到位

☑ 避免：只是描述现象，不分析原因，不回应科学假设

如何写作？ —— Results/Experiments

起：

4.1 Setup

Office-31 is a standard benchmark for visual domain adaptation, which contains 4,652 images and 31 categories from three distinct domains: Amazon (A), Webcam (W) and DSLR (D).

Office-Home is a more challenging domain adaptation dataset than Office-31, which consists of around 15,500 images from 65 categories of everyday objects. This dataset is organized into four domains: Art (Ar), Clipart (Cl), Product (Pr) and Real-world (Rw).

Compared Approaches

Beside the classical approaches, we also compare our disentangled semantic representation model with some deep transfer learning methods. Two recently proposed semantic enhanced methods, CDAN [Long *et al.*, 2018] and MSTN [Xie *et al.*, 2018], are also compared in the experiment. Note that, CDAN conditions the adversarial adaptation models on discriminative information conveyed in the classifier predictions, and MSTN learns semantic representation by aligning labeled source centroid and pseudo-labeled target centroid.

分：

Office-home Result

The Study of the Disentangled Semantic Representation

Ablation Study of the Dual Adversarial Learning

如何写作？——Conclusion

☑ 目的：总结科学假设验证的意义，阐明take away的东西

☑ Style

☑ 力求：总结要恰当，意义要深刻

☑ 避免：成为abstract的翻版

如何写作？——Conclusion

总结：

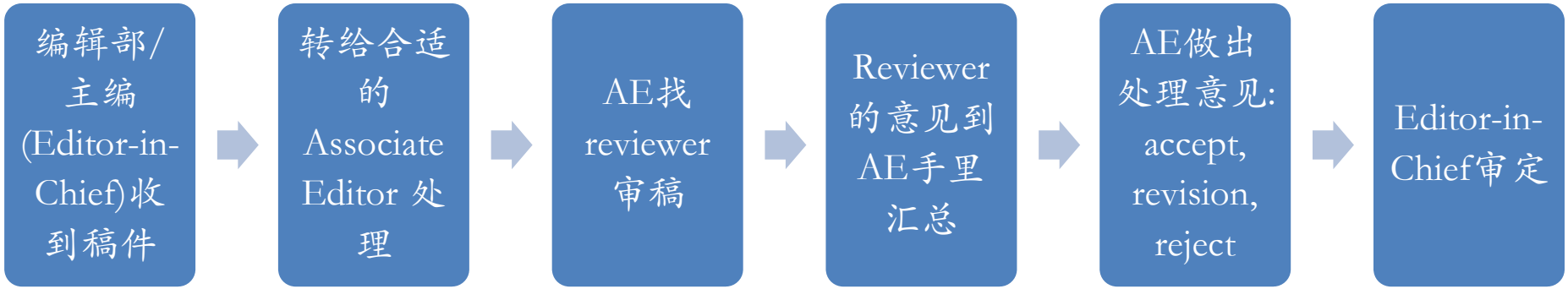
This paper presents a disentangled semantic representation model for the unsupervised domain adaptation task. Different from previous work , our approach extracts the disentangled semantic representation on the recovered latent space, following the causal model of the data generation process. Our approach is also featured with the variational auto-encoder based latent space recovery and the dual adversarial learning based disentangle of the representation.

意义：

The success of the proposed approach not only provides an effective solution for the domain adaptation task, but also opens the possibility of disentanglement based learning methods.

如何写作？—— after submission

☑ 稿件处理流程：杂志为例



如何写作？—— after submission

☑ 目的：Revision，解释不清楚的点，提升文章质量

☑ Style

Comment 1.2： *More context and general problem statement are needed inside the introduction*

Reply: We have revised the introduction section by providing more context and problem statement. Thanks for your suggestion.

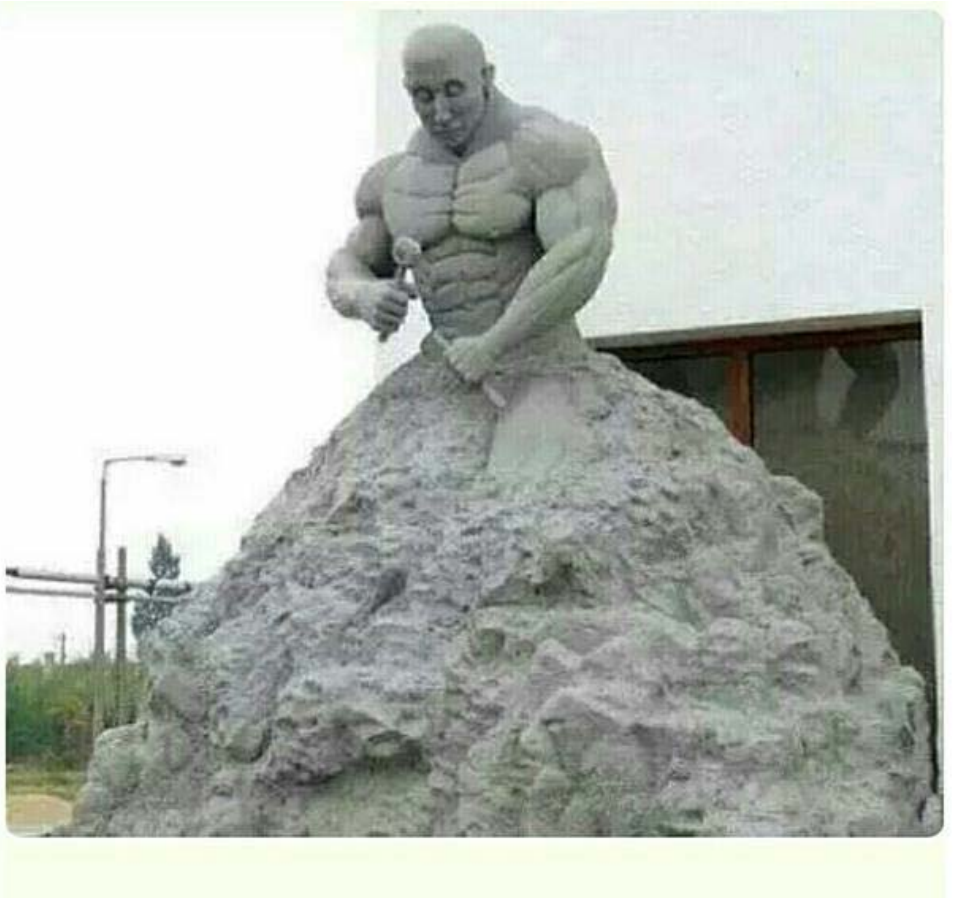
☑ 力求：理解审稿人意思，尽量满足审稿人要求

☑ 避免：强词夺理

如何写作？ —— 一些误区

☑ 误区1：The more, the better

☑ 原则：抓住“core work”，围绕着core work准备的材料越充分，越好。不是core work尽量不涉及



如何写作？ —— 一些误区

- ☑ 误区2：The more complex, the better

- ☑ 原则1：用户理解你的工作是第一位要务。

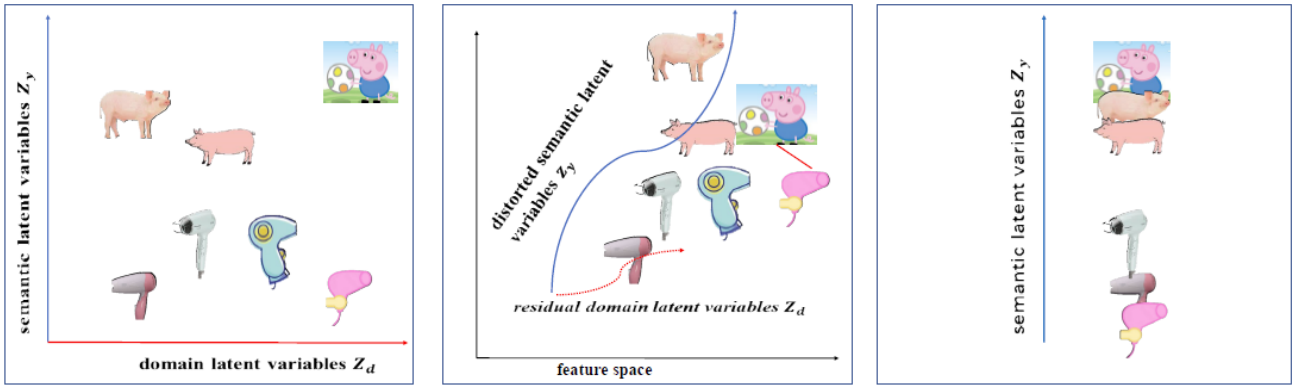
- ☑ 原则2：公式与定理是确保工作的严密性，而非炫耀

- ☑ 原则3：公式与定理需要文字来解释

如何写作？ —— 一些体会

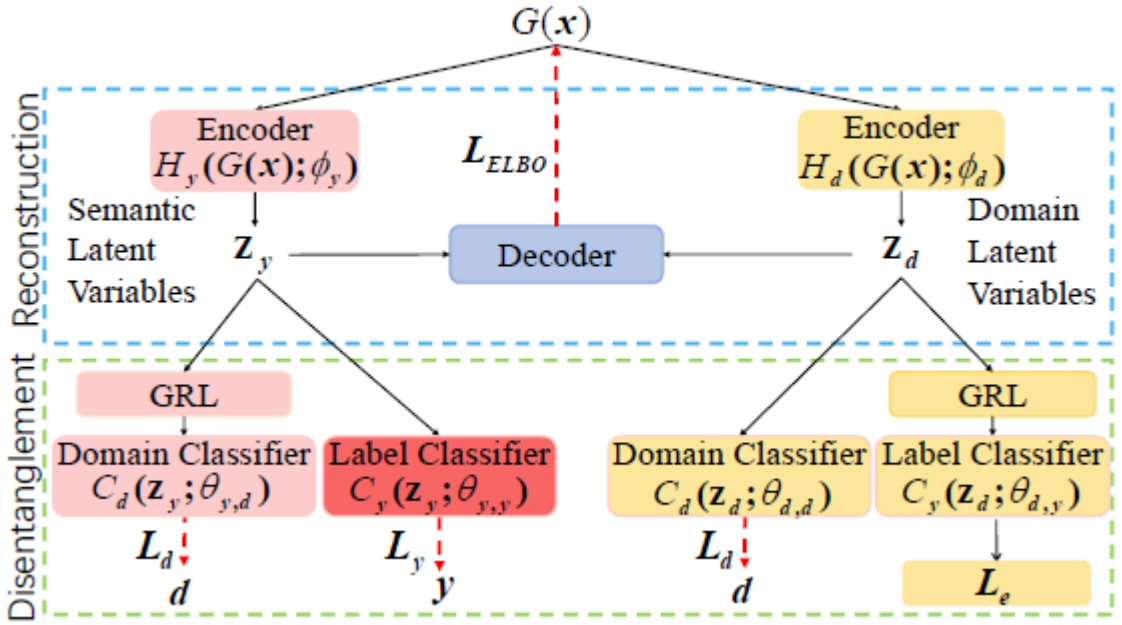
☑ 一图胜万言

- Motivation 图



(a) Latent manifold of the data generation. (b) Distorted feature manifold with residual (c) Disentangle semantic manifold

- 方案框架图



如何写作？ —— 一些体会

☑ 经验1： 多读多写多总结

- 无他，唯手熟耳
- 形成自己的方法论

☑ 经验2：找个论文作为模板

- 体会、模仿而非抄袭

☑ 经验3：接受不确定性

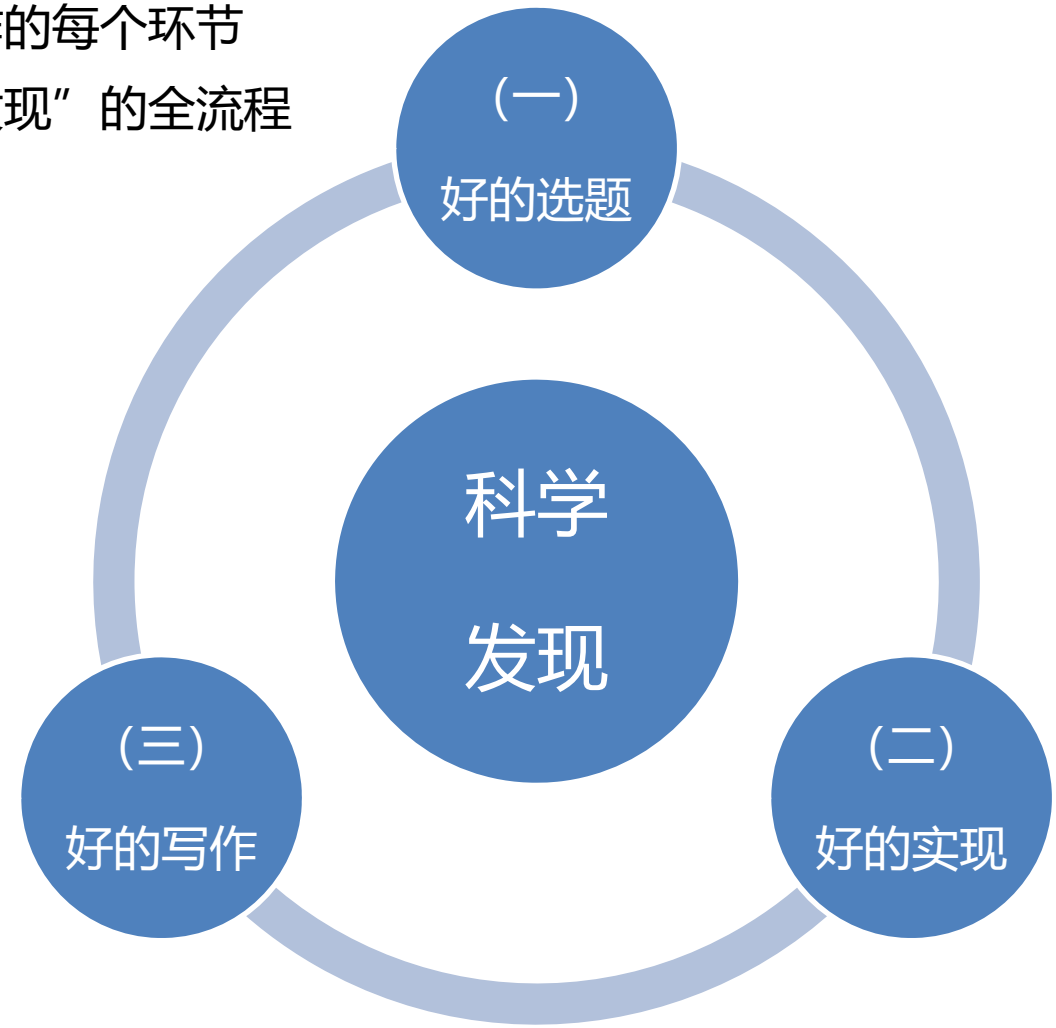
- 众口难调，有所坚持
- 斐波那契投稿法：本次投稿=上次被拒+上上次被拒

☑ 经验4：逻辑，逻辑，逻辑！！！

小结

小结

- ☑ 回到初心
- ☑ 做好选题、研究、写作的每个环节
- ☑ 完美演绎“验证科学发现”的全流程

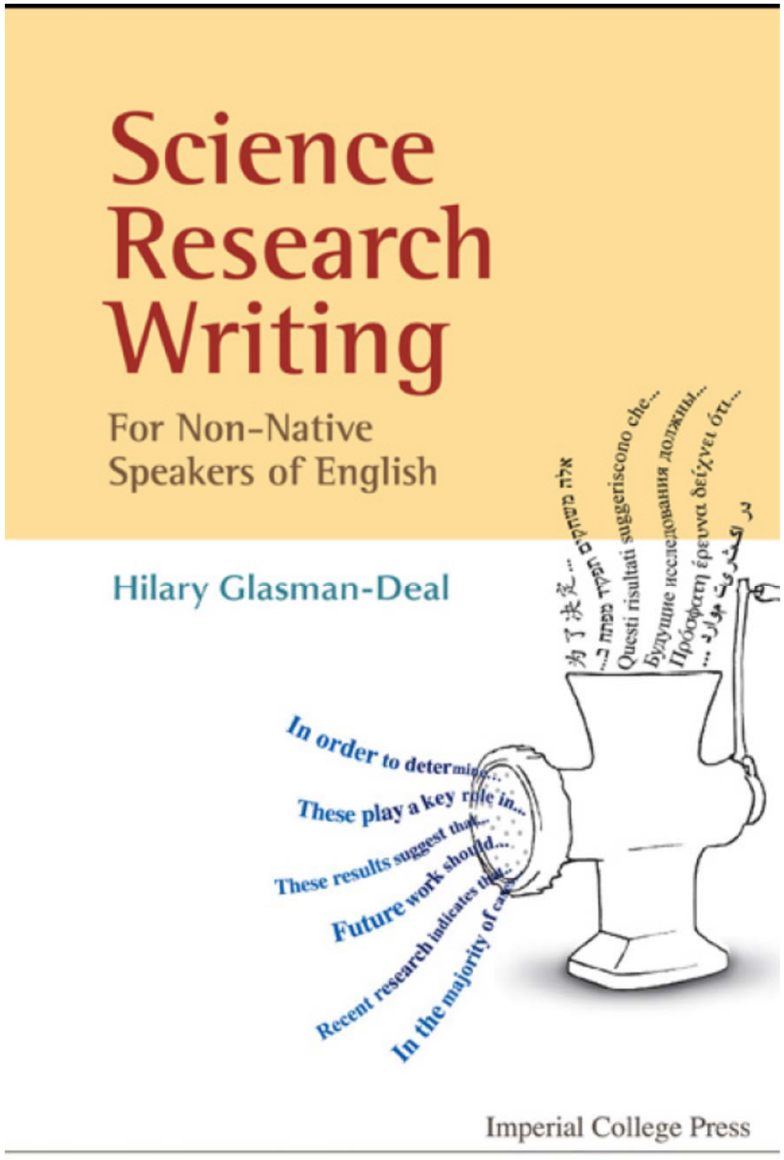


小结

文章千古事，得失寸心知

清风明月

参考


Science
Research
Writing
For Non-Native
Speakers of English
Hilary Glasman-Deal
In order to determine...
These play a key role in...
These results suggest that...
Future work should...
Recent research indicates that...
In the majority of cases...
Imperial College Press


麦肯锡三十年
经典培训教材
金字塔
原理
The Minto
Pyramid Principle

Thank you!

http://www.dmirlab.com/

CAMAL 2019

International Workshop on Casual Modeling and Machine Learning

Xiaohua Zhou
AAAS Fellow
Peking University

Lei Xu
EURASC Fellow
Shanghai Jiao Tong University

Jiji Zhang
Associate Prof.
Lingnan University

Kun Zhang
Assistant Prof.
Carnegie Mellon University

Mingming Gong
Assistant Prof.
University of Melbourne

Ruichu Cai
Professor
Guangdong University of Technology

Sara Magliacane
Researcher
MIT-IBM Watson AI lab in Cambridge

Shengyu Zhu
Researcher
Huawei Noah's Ark lab

Tom Claassen
Researcher
Radboud University Nijmegen

Yangbo He
Associate Prof.
Peking University

本次研讨会将于2019年11月23日在中国广州大学城举办，其旨在为对因果关系和机器学习感兴趣的不同研究领域的研究学者和研究生提供一个交流的平台。研讨会包括特邀报告，海报展示以及足够的时间的科学讨论。本次研讨会的主题包括但不限于：

- 因果关系的特征形式化
- 从因果视角解决机器学习问题和算法
- 智能系统的因果关系
- 因果发现和推断的计算方法
- 因果发现和推断的实际应用
- 心理学和神经科学的因果关系