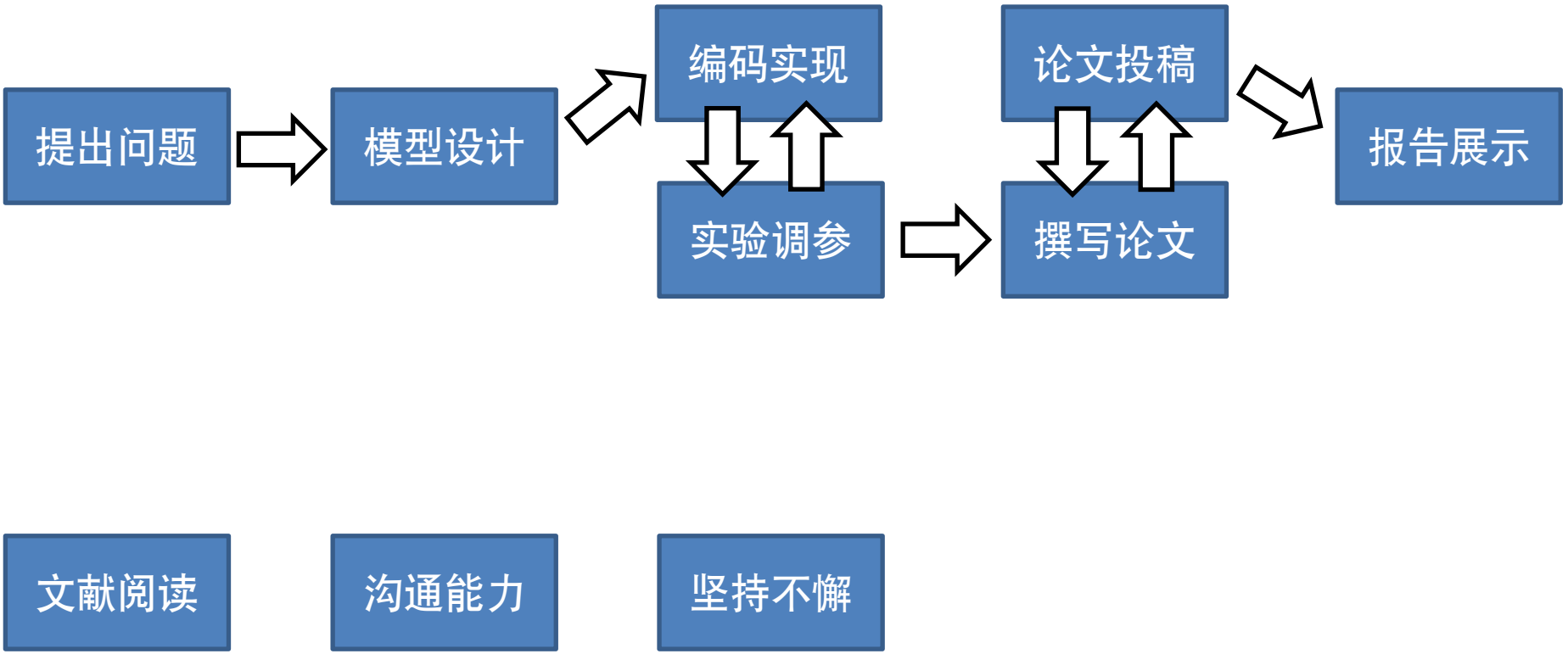
CCL 2018 学生研讨会

文献综述与研究选题

清华大学自然语言处理实验室

刘知远

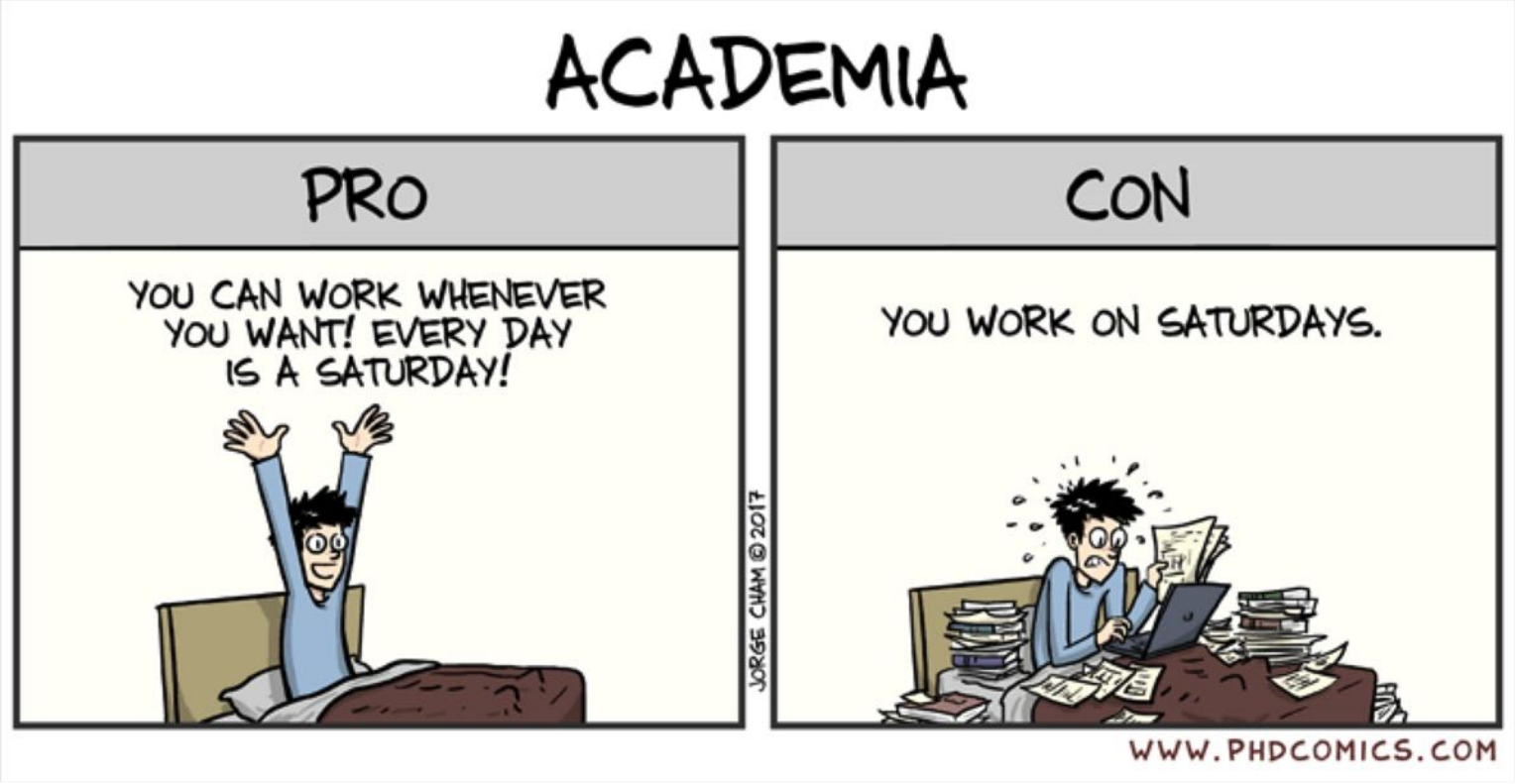
学术研究是一项系统工程



学术研究需要天时地利人和

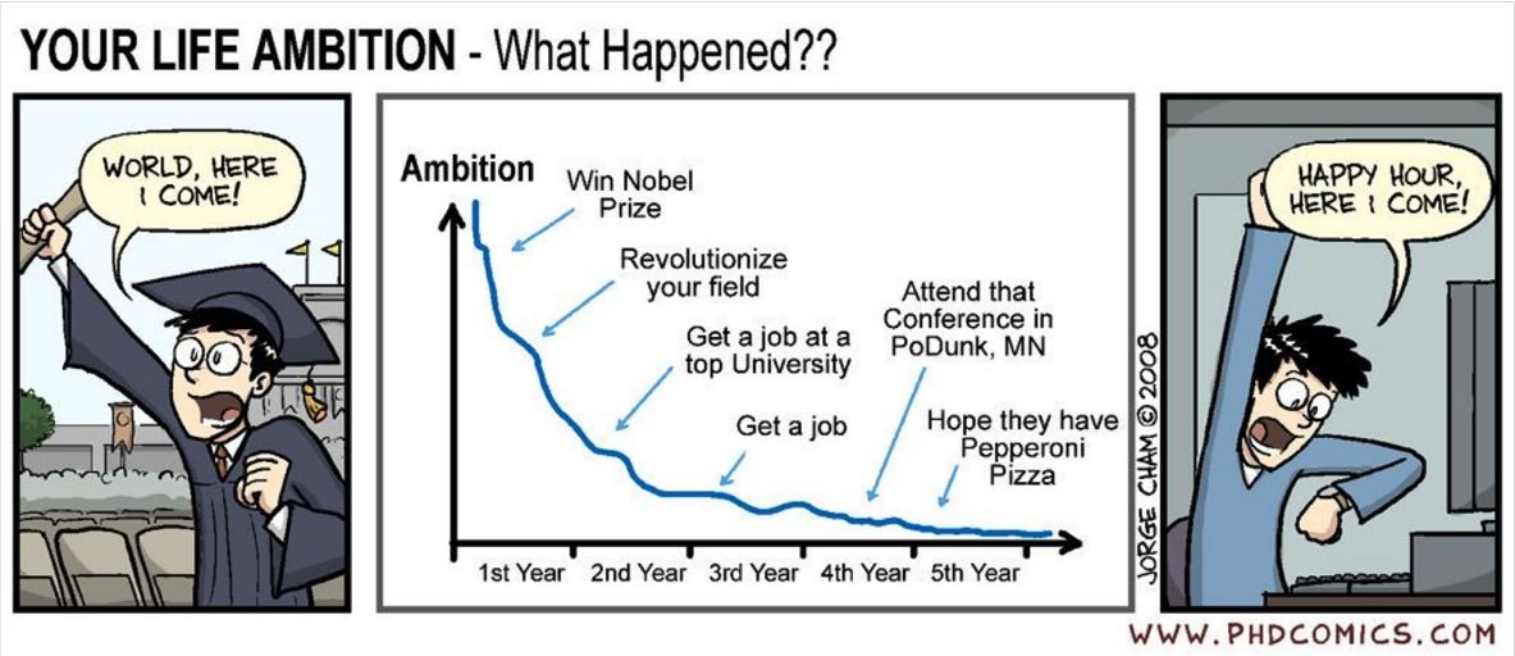
成功的研究 =

重要问题 + 新颖方法 + 努力、积累、坚持

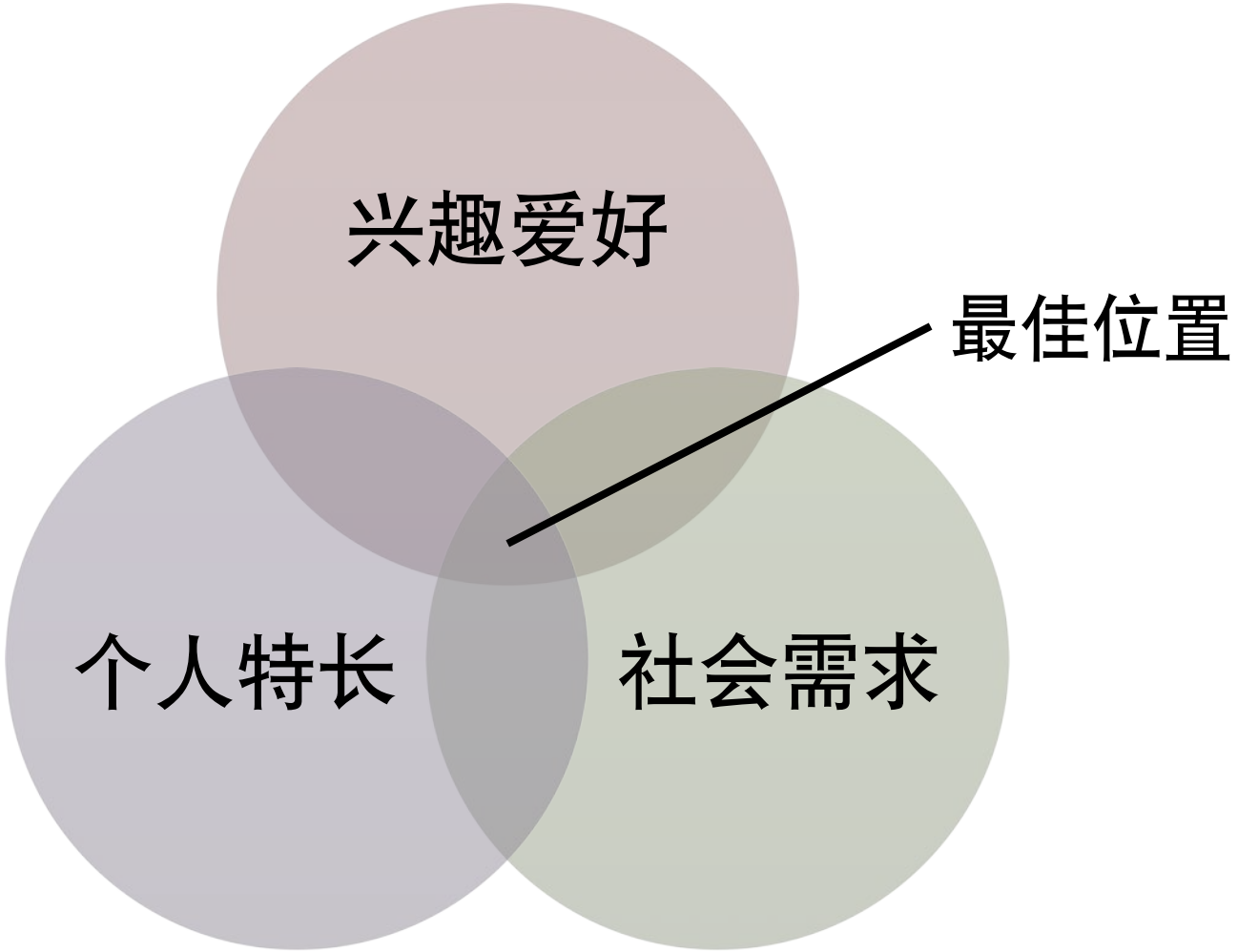


学术研究不同时期有不同追求

- 第一层：锻炼解决开放问题的能力
- 第二层：成为相关领域的知名专家
- 第三层：做出引领领域方向的工作



研究方向的选择

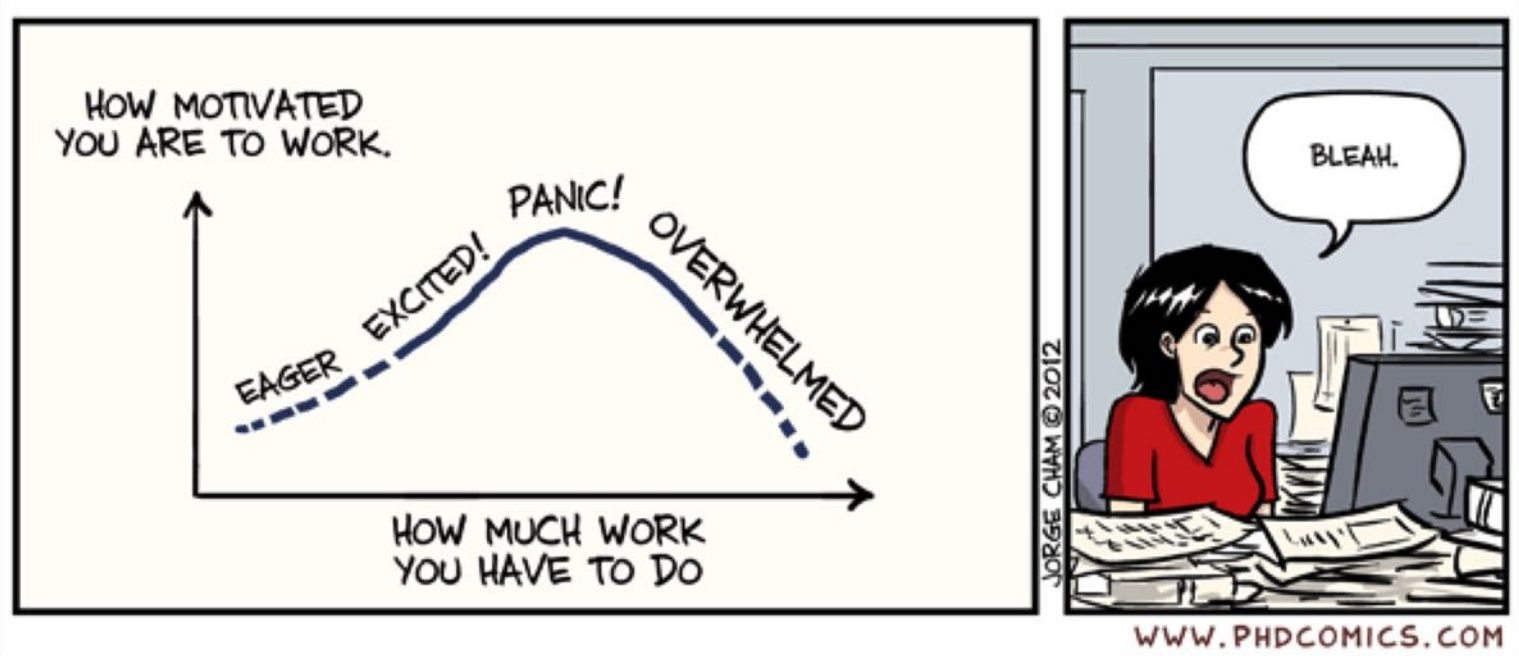


研究建议-1

- 正视个体差异
- 扬长避短：不同研究环节侧重不同方面
- 循序渐进：
  - 从事第一项研究时，可主要负责模型具体设计与实现，导师/学长主要负责选题、技术路线设计和论文撰写
  - 成功完成首项研究任务后，则可以开始在选题等方面承担更多责任，从而得到更全面的锻炼

研究建议-2

- 迅速进入研究实战状态
  - 在学习入门知识的同时，迅速从具体研究任务入手，开始研究历练
  - 在实践中学习，学以致用，实现对领域的全景式了解

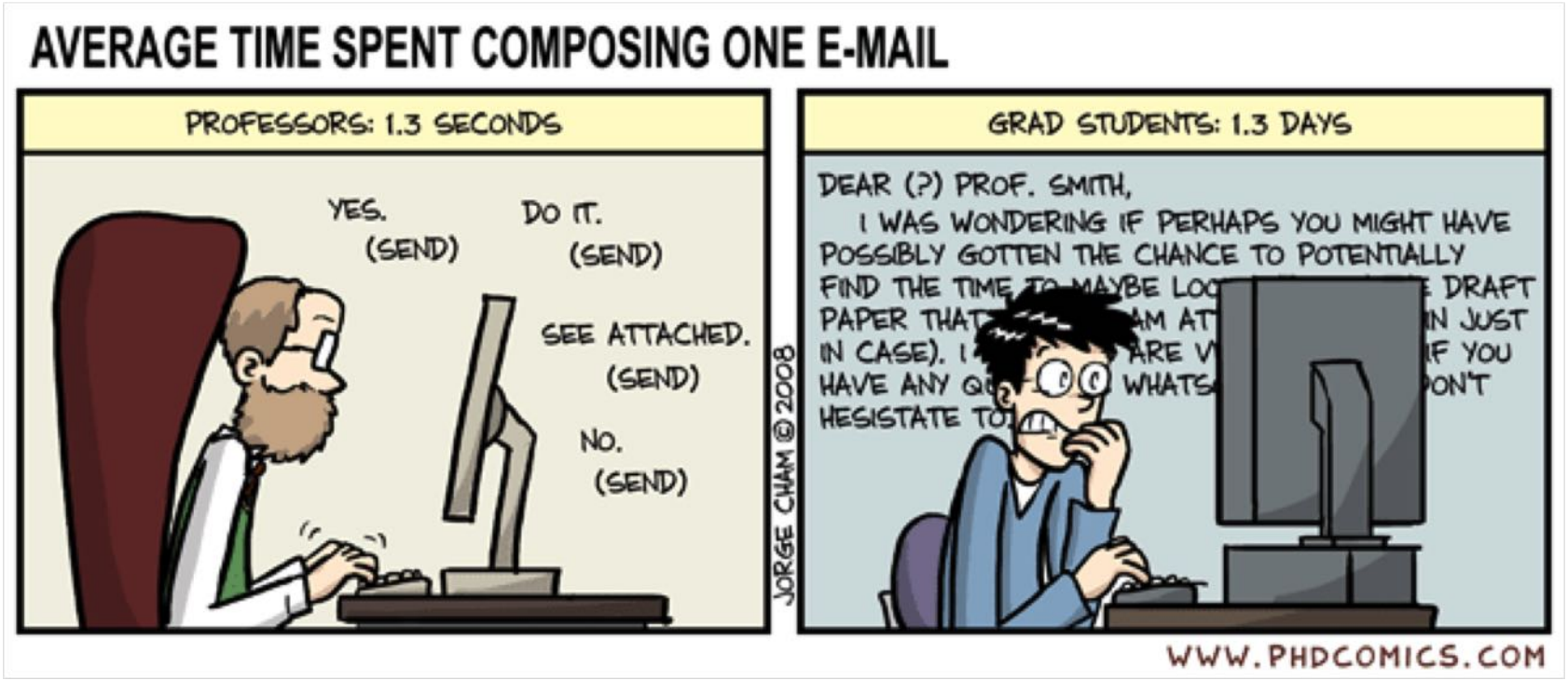


研究建议-3

- 将科研列为高优先级
  - 无数事例表明，研究成绩与重视程度成正比
  - 正式加入实验室前，慎重决定，一旦决定全力以赴

研究建议-4

- 坚持主动积极的态度
  - 积极与导师学长交流，充分利用Lab资源，一切以完成高水平研究为目标



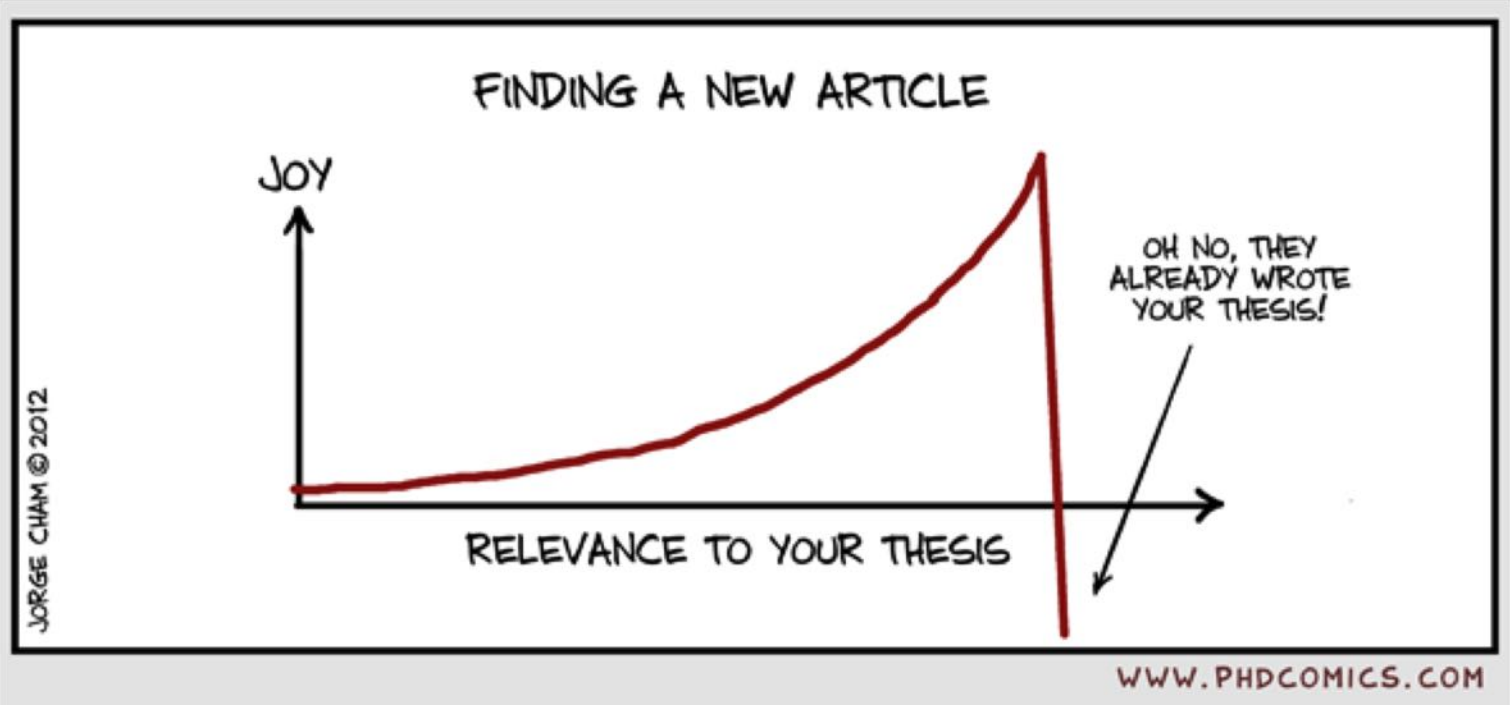
如何查阅文献

如何查找论文（给定关键词）

维基百科
（搜索引擎）

中文综述
（CNKI）

英文论文
（Google Scholar）

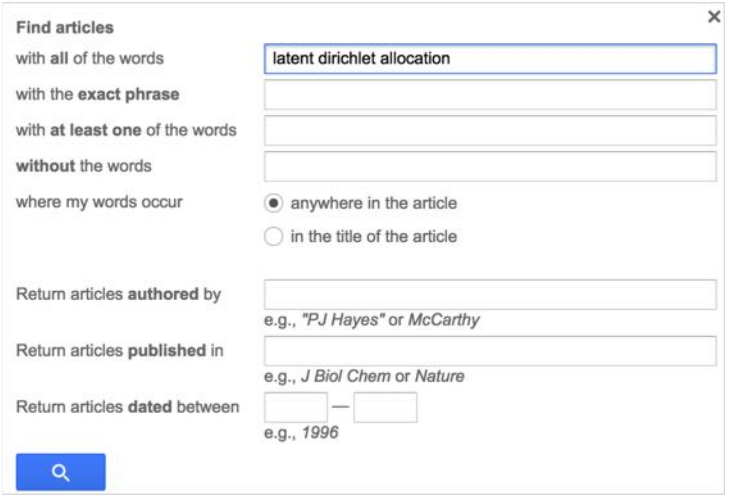


善用Google Scholar

- 查阅学者学术信息、引用情况，也提供引用格式文件

Latent dirichlet allocation
DM Blei, AY Ng, MI Jordan - Journal of machine Learning research, 2003 - jmlr.org
Abstract We describe latent Dirichlet allocation (LDA), a generative probabilistic model for collections of discrete data such as text corpora. LDA is a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying ...
Cited by 15978 Related articles All 124 versions Import into BibTeX Cite Save Fewer

- 学会使用相关搜索命令
  - Author: “DM Blei”
  - AllInTitle:“Latent dirichlet allocation”
  - …



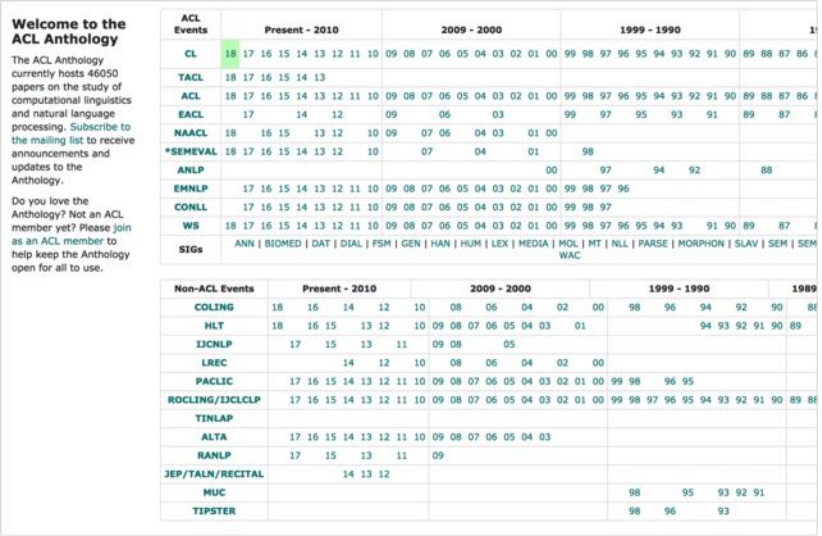
如何判断论文是否值得阅读

- 作者是否大牛学者？作者机构是否顶尖？
- 是否发表在顶级期刊/会议上？
- 论文社会关注度如何？是否获得最佳论文？引用情况如何？

学术资源-ACM

- 美国计算机学会
- 全球最大的计算机学术组织
- ACM DI拥有大量高水平论文
  - 信息检索
  - 数据挖掘
  - …

学术资源-ACL



Welcome to the ACL Anthology

The ACL Anthology currently hosts 46050 papers on the study of computational linguistics and natural language processing. Subscribe to the mailing list to receive announcements and updates to the Anthology.

Do you love the Anthology? Not an ACL member yet? Please join as an ACL member to help keep the Anthology open for all to use.

ACL Events	Present - 2010	2009 - 2000	1999 - 1990	19
CL	18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86
TACL	18 17 16 15 14 13			
ACL	18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 88 87 86
EACL	17 14 12	09 06 03	99 97 95 93 91	89 87
NAACL	18 16 15 13 12 10	09 07 06 04 03 01 00		
*SEMEVAL	18 17 16 15 14 13 12 10	07 04 01	98	
ANLP		00	97 94 92	88
EMNLP	17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96	
CONLL	17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97	
WS	18 17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 91 90	89 87
SIGs	ANN | BIOMED | DAT | DIAL | FSM | GEN | HAN | HUM | LEX | MEDIA | MOL | MT | NLL | PARSE | MORPHON | SLAV | SEM | SEM WAC			

Non-ACL Events	Present - 2010	2009 - 2000	1999 - 1990	1989
COLING	18 16 14 12 10	08 06 04 02 00	98 96 94 92 90	8
HLT	18 16 15 13 12 10	09 08 07 06 05 04 03 01	94 93 92 91 90	89
IJCNLP	17 15 13 11	09 08 05		
LREC	14 12 10	08 06 04 02 00		
PACLIC	17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 96 95	
ROCLING/IJCLCLP	17 16 15 14 13 12 11 10	09 08 07 06 05 04 03 02 01 00	99 98 97 96 95 94 93 92 91 90	89 8
TINLAP				
ALTA	17 16 15 14 13 12 11 10	09 08 07 06 05 04 03		
RANLP	17 15 13 11	09		
JEP/TALN/RECITAL	14 13 12			
MUC			98 95 93 92 91	
TIPSTER			98 96 93	

- 国际计算机学会
- 全球最大的自然语言处理学术组织
- ACL Anthology囊括几乎全部的NLP重要论文（全部免费）
  - ACL
  - NAACL
  - EMNLP
  - COLING
  - …

学术资源-CCL

http://www.cips-cl.org/anthology

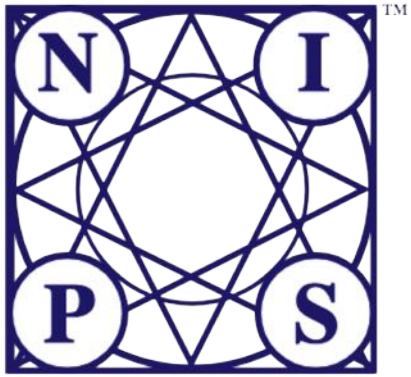
标题	作者	会议	下载量
基于深度学习的微博情感分析	梁军，柴玉梅，原慧斌，昝红英，刘铭	CCL 2014	2747
基于表示学习的中文分词算法探索	来斯惟，徐立恒，陈玉博，刘康，赵军	CCL 2013	1846
基于深度学习加强的混合推荐方法	丁弼原，张敏，谭云志，刘奕群，马少平	CCL 2016	1699
基于卷积神经网络的微博情感倾向性分析	刘龙飞，杨亮，张绍武，林鸿飞	CCL 2015	1690
结合卷积神经网络和词语情感序列特征的中文情感分析	陈钊，徐睿峰，桂林，陆勤	CCL 2015	1677
基于极性转移和LSTM递归网络的情感分析	梁军，柴玉梅，原慧斌，高明磊，昝红英	CCL 2015	1456
基于评论挖掘的药物副作用发现机制	程亮喜，赵明珍，林鸿飞	CCL 2014	1425
基于句法语义特征的中文实体关系抽取	郭喜跃，何婷婷，胡小华，陈前军	CCL 2014	1337
基于规则的越南语命名实体识别研究	闫丹辉，毕玉德	CCL 2014	1306
知识图谱中实体相似度计算研究	李阳	CCL 2016	1211

论文下载量排行榜

统计时间：2016年3月至2018年9月

学术资源-ICML/NIPS

- 机器学习领域的两大顶级会议



- 深度学习时代的新兴学术会议

学术资源-Arxiv

arXiv.org

- 预印本文库
- 未发表的论文，良莠不齐
- 建议关注顶级组织的相关论文

subscribe Zhiyuan Liu

1 message

Zhiyuan Liu <liuzy@tsinghua.edu.cn>
To: cs@arxiv.org

add CL
add LG
add NE

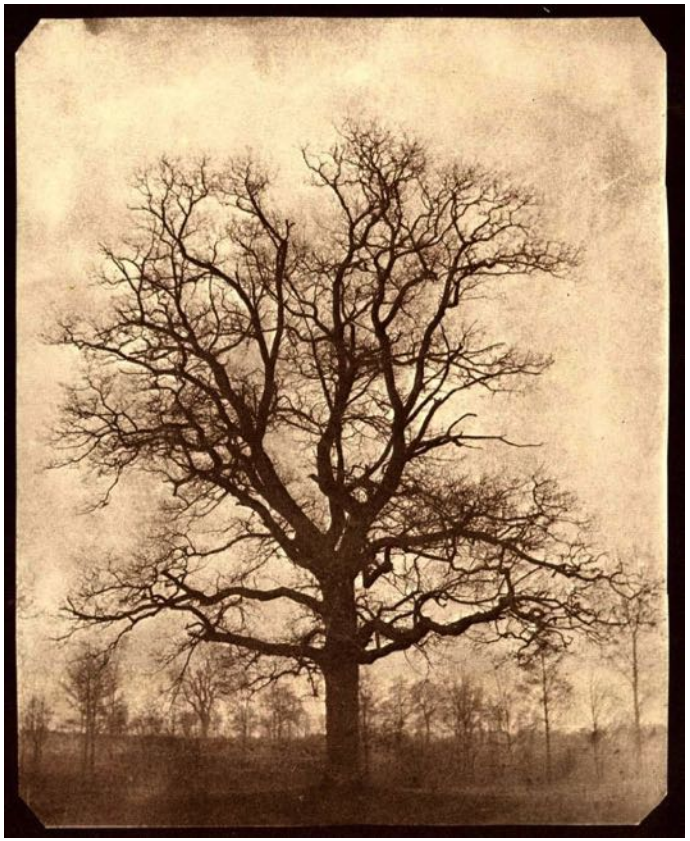
阅读论文顺序

- 题目（1）
- 摘要（2）
- 正文
  - 导论（3）、相关工作、自己工作（5）、实验结果（4）、结论
- 致谢
- 参考文献（6）
- 附录

如何研究选题

提出问题比解决问题更重要

- 一流学者提出问题
- 二流学者解决问题
- 三流学者打补丁



本页来自清华大学刘洋老师《撰写学术论文的若干技巧》。

为什么找问题更重要、更难？

- 提出问题者往往能影响整个领域的发展方向
- 解决问题往往是个技术活，能够后天培养（理论素养、编程能力、写作能力等），而提出问题则需要：
  - 站得更高
  - 看得更远
  - 嗅觉更好
  - 当机立断
  - 不畏风险

本页来自清华大学刘洋老师《撰写学术论文的若干技巧》。

如何找问题？

Think differently

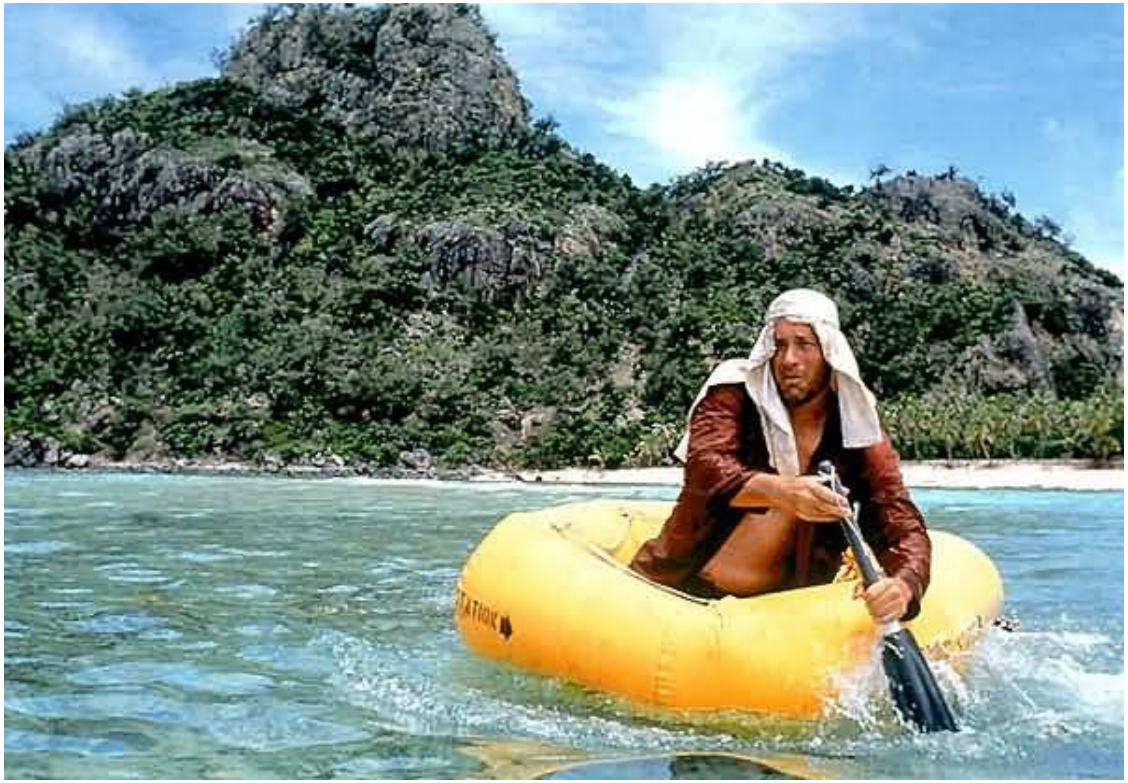
满腹经纶者固然可敬，擅长推陈出新者更值得推崇。

本页来自清华大学刘洋老师《撰写学术论文的若干技巧》。

哪里热闹去哪里

本页来自清华大学刘洋老师《撰写学术论文的若干技巧》。

哪里人少去哪里



*“It is not worth an intelligent man's time to be in the majority. **By definition**, there are already enough people to do that.”*

--- G. H. Hardy (1877-1947)

本页来自清华大学刘洋老师《撰写学术论文的若干技巧》。

如何找到好问题

- 博览群书，对整个领域有全貌式把握
- 熟知学术界动态，知道当前最热门问题是什么
- 明察秋毫，富有远见，结合个人兴趣选择一个数年后会变成热门的领域，并全力以赴去做

本页来自清华大学刘洋老师《撰写学术论文的若干技巧》。

做好不被认可的准备

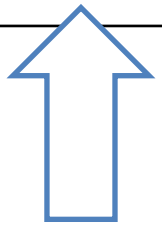
Frank Rosenblatt
1928–1969

Rosenblatt's perceptron played an important role in the history of machine learning. Initially, Rosenblatt simulated the perceptron on an IBM 704 computer at Cornell in 1957, but by the early 1960s he had built special-purpose hardware that provided a direct, parallel implementation of perceptron learning. Many of his ideas were encapsulated in "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms" published in 1962. Rosenblatt's work was criticized by Marvin Minksy, whose objections were published in the book "Perceptrons", co-authored with Seymour Papert. This book was widely misinterpreted at the time as showing that neural networks were fatally flawed and could only learn solutions for linearly separable problems. In fact, it only proved such limitations in the case of single-layer networks such as the perceptron and merely conjectured (incorrectly) that they applied to more general network models. Unfortunately, however, this book contributed to the substantial decline in research funding for neural computing, a situation that was not reversed until the mid-1980s. Today, there are many hundreds, if not thousands, of applications of neural networks in widespread use, with examples in areas such as handwriting recognition and information retrieval being used routinely by millions of people.

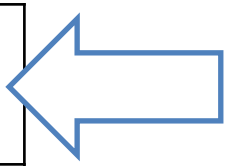
Pattern Recognition and Machine Learning, C. Bishop

科研选题中的创新问题

	老方法	新方法
老问题	✗	✓
新问题	✓	✓



语言学相关领域文献

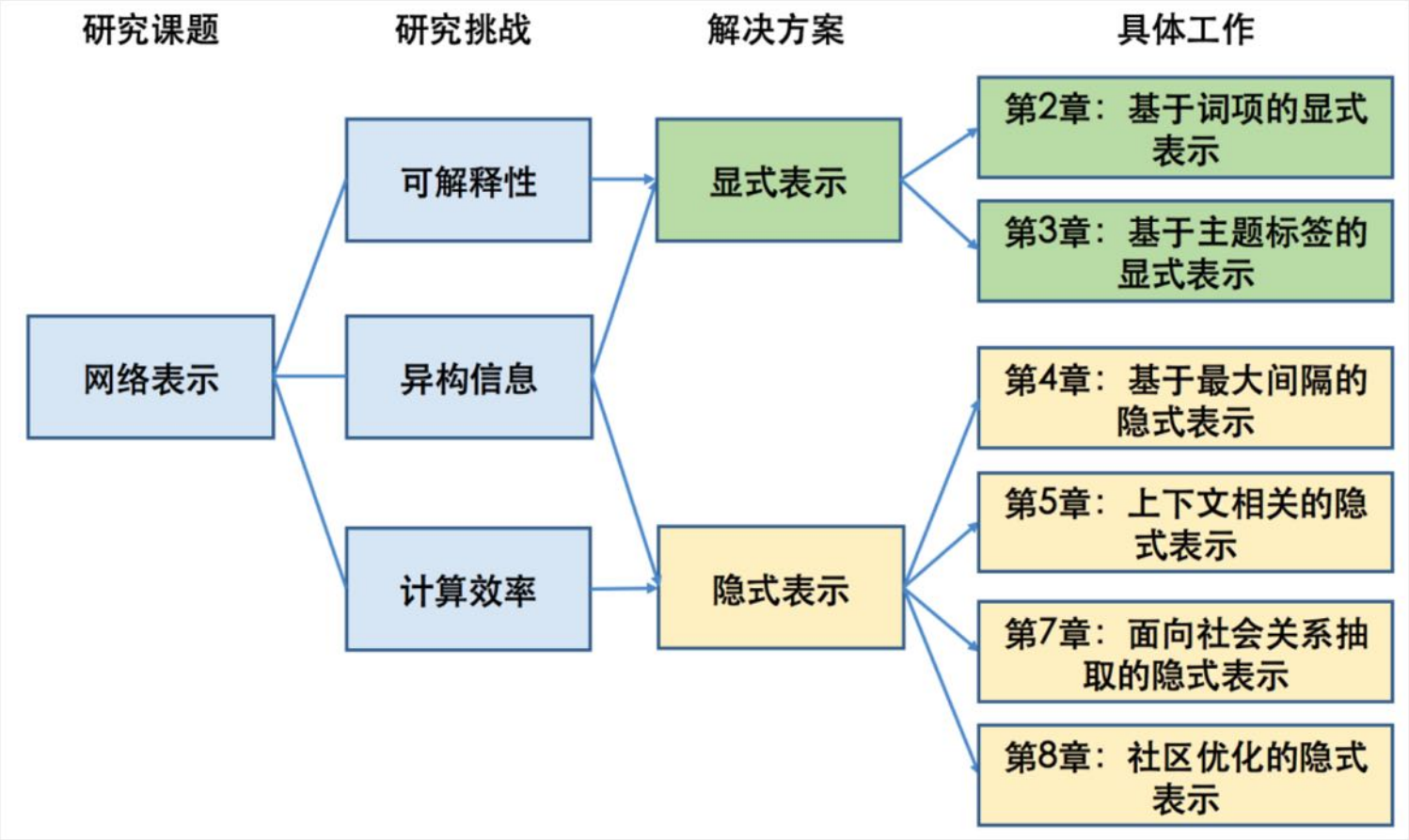


机器学习相关领域文献

对博士生的选题建议

- 不只训练对单独一份工作选题能力
- 思考博士生涯的整体选题（3-5个独立工作）

对博士生的选题建议



涂存超（2018）：面向社会计算的网络表示学习

全国NLPers，联合起来！

http://nlp.csai.tsinghua.edu.cn/~lzy/

liuzy@tsinghua.edu.cn