从数据中学习

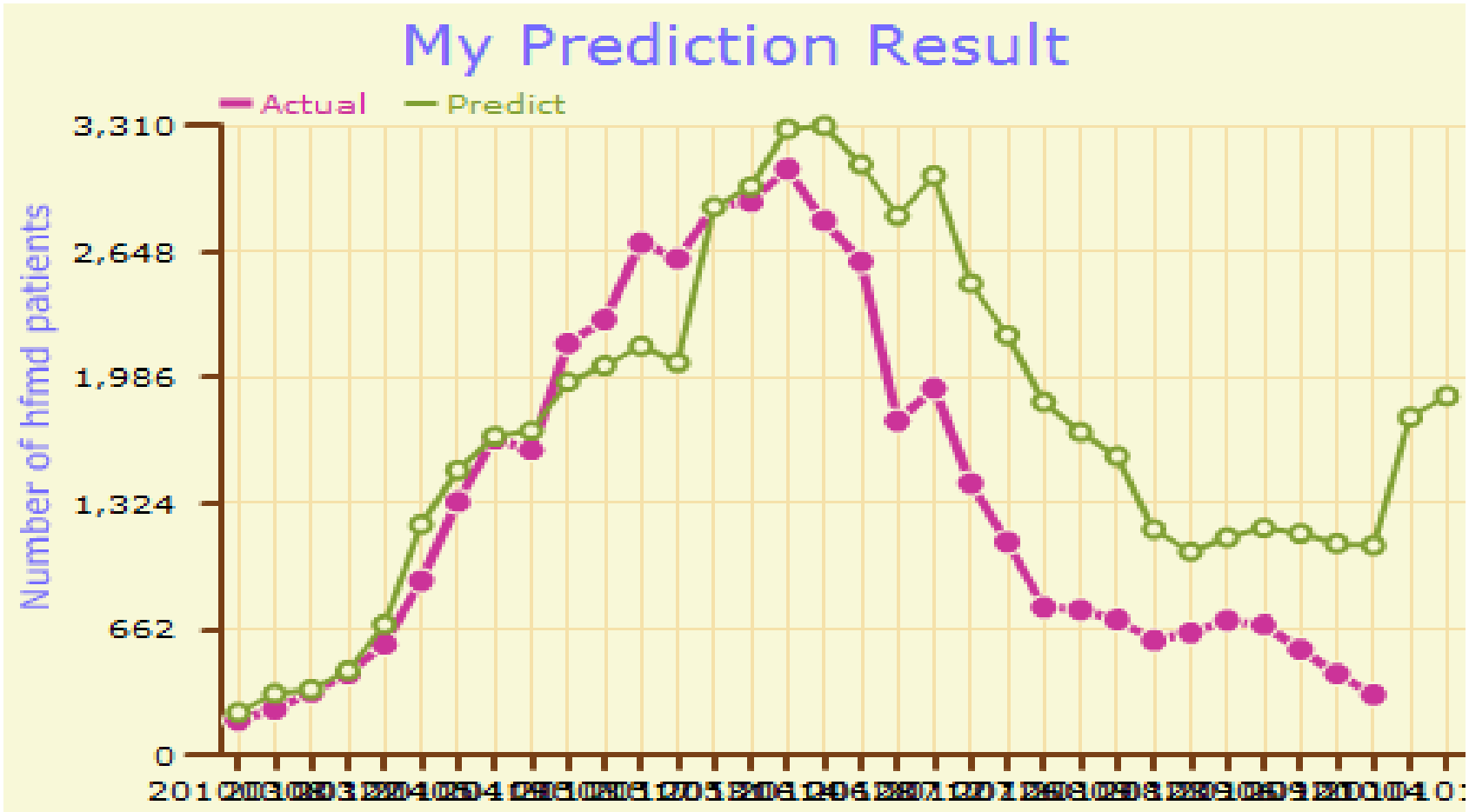
马少平 清华大学
智能技术与系统国家重点实验室

海量数据的威力

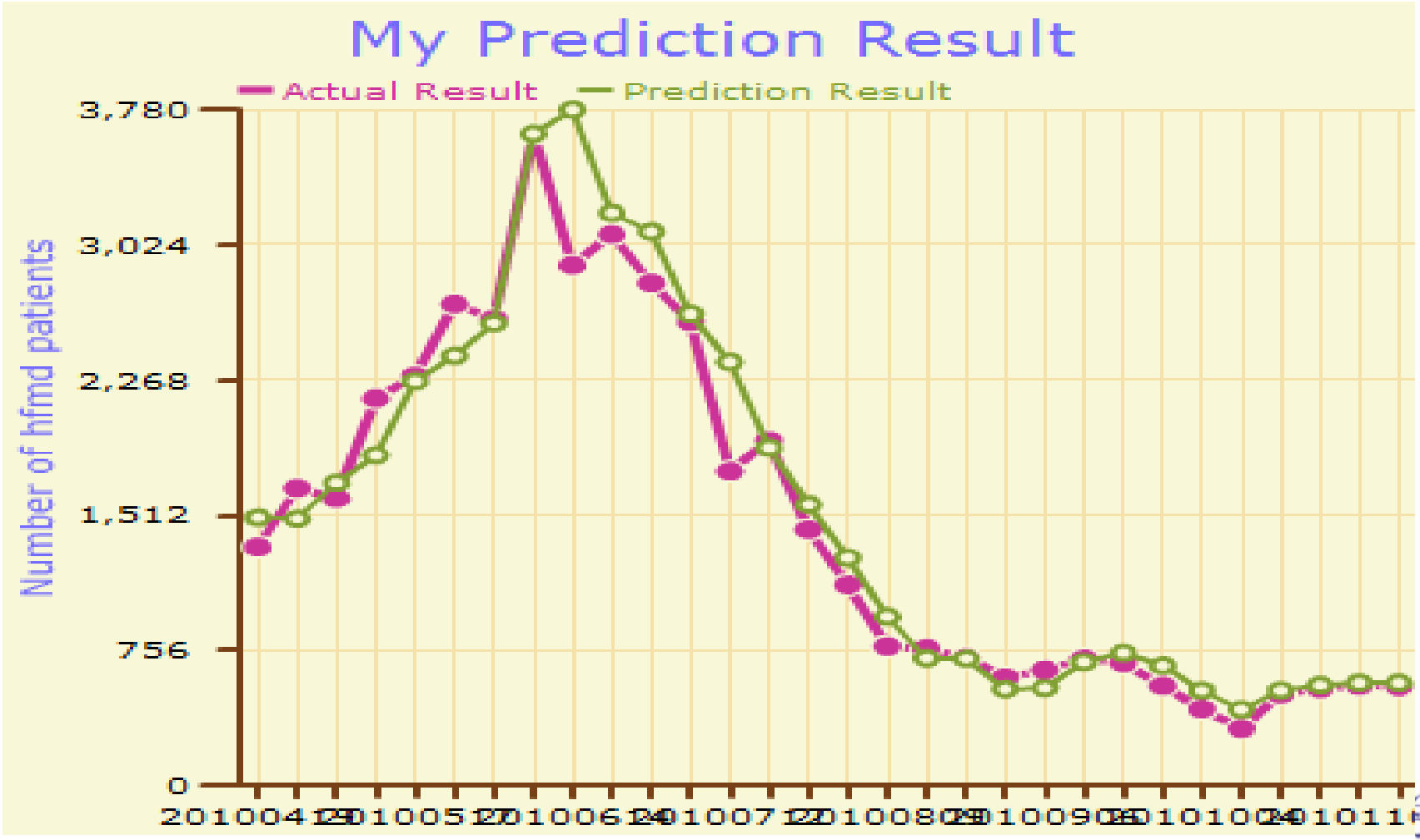
- 数据量大
- 种类多
- 知识蕴含其中
- 知识的获得反而简单
- 现有方法已经可以有所作为
- 既是挑战，又是机会

从皇后问题想到的

流行病预测


My Prediction Result
Actual
Predict
Number of hfmd patients
3,310
2,648
1,986
1,324
662
0

流行病预测


My Prediction Result
Actual Result
Prediction Result
Number of hfmd patients
3,780
3,024
2,268
1,512
756
0

查询纠错


新闻 网页 音乐 图片 视频 地图 知识 更多>>
Sogou搜狗
sunmaosong
搜狗搜索
找到 23,500 个网页，用时 0.162 秒
网页结果
新 闻
音 乐
图 片
视 频
微 博
问 答
百 科
更 多
全部时间
一天内
智能提示 - 您是不是要找：孙茂松
孙茂松_百度百科
清华大学教授、博士生导师——孙茂松 人物简介 孙茂松，清华大学计算机科学与技术系系主任，教授，博士生导师。研究方向为自然语言理解、中文信息处理和Web智能。作为项...
baike.baidu.com/view/312782.htm - 2012-4-9 - 快照
实验室人员信息
孙茂松 Sun Maosong孙茂松，清华大学计算机科学与技术系副系主任，教授，博士生导师。研究领域为计算语言学、中文信息处理、信息检索和人工智能。作为项目负责人，主持...
www.csai.tsinghua.edu.cn/...ab/sunmaosong.shtm - 2011-10-13 - 快照

微博信息处理

- 孙茂松 **32.3011**
- 刘挺 **29.0042**
- 王斌_ICTIR **29.001**
- 李航博士 **26.201**
- 梁斌penny **26.1015**
- 微软亚洲研究院 **24.901**
- 刘知远THU **23.9003**
- 张敏THU **23.201**
- 白硕sse **21.1045**
- 王海峰_百度 **20.301**

机器学习 ok

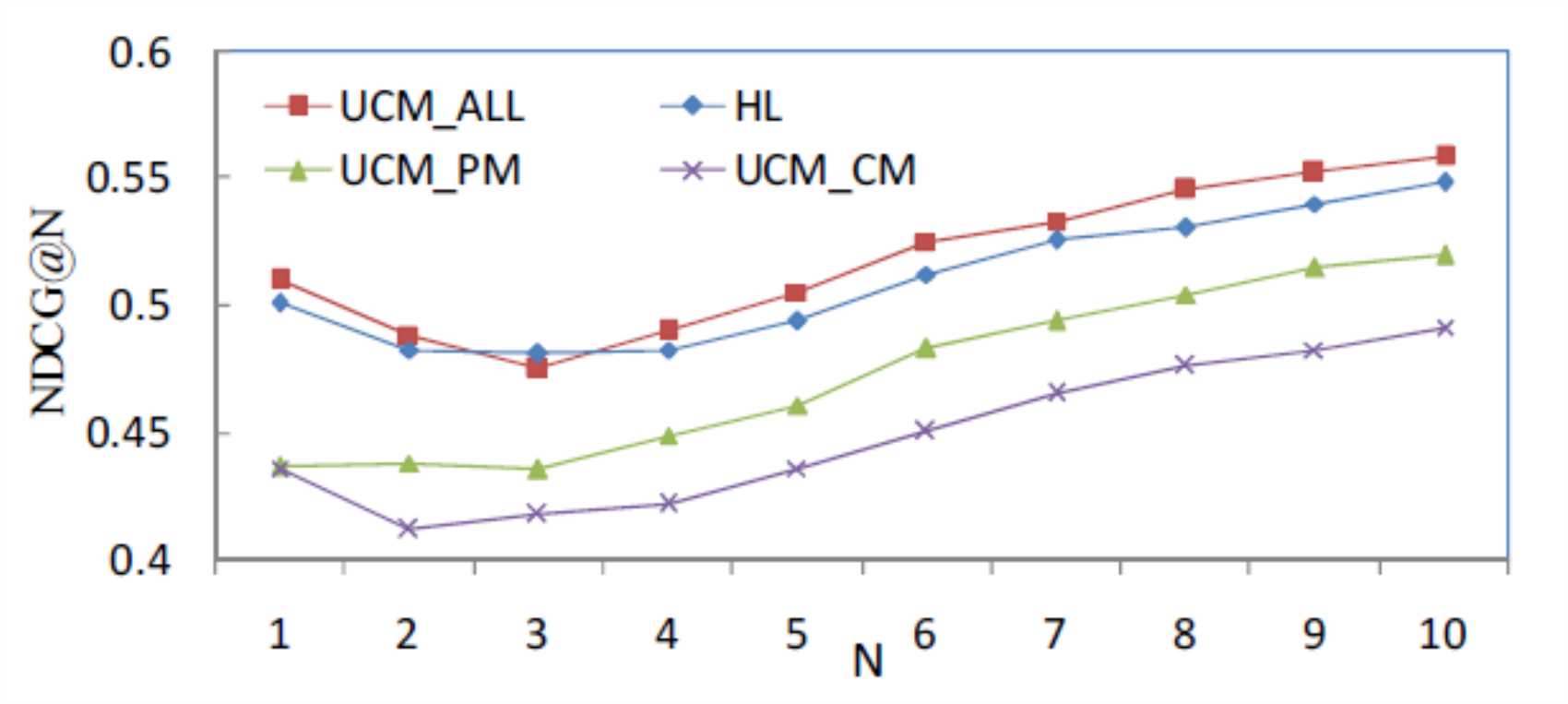
weibo_cnt	att_cnt	fans_cnt	pic	name
284	95	8968		#李航博士#
999	290	13903		#张栋_机器学习#
311	127	1835		#孙茂松#
531	295	926		#张夏天_机器学习#
5166	1990	1606		#丕子#
373	212	461		#ITNLP#
141	92	619		#Chen_1st#
57	179	646		#许冬亮_感知世界#
144	139	6423		#胡云华MSRA#
282	258	918		#xlvector_Hulu#
254	147	294		#朱小飞_计算所#

属性词、观点词挖掘


机身重量轻，做工不错，电池也比较耐用。只是感觉塑料味儿有些浓！
Its weight is light, quality is good, battery is durable. But Its flavor of plastic is heavy!
FW: Feature Words
OW: Opinion Words
S: sentiment polarity

- [1] 按键布局|合理 Button Layout | Reasonable
- [1] 按键布局|简洁 Button Layout | Concise
- [1] 镜头 | 给力 Camera lens | Geili (pretty good)
- [1] 按键手感|舒适 Button Touch Feel | Comfortable
- [-1] 暗部细节|欠缺 Dark Details | Lack of
- [-1] 白天色彩|暗淡 Color in Daylight | Dim
- [-1] 变焦杆|硬 Zoom Lever | Stiff
- [1] 材质|耐磨 Material | Abrasion Resistant
- [1] 长焦|强悍 Telephoto lens | Extreme
- [-1] 价格 | 高 Price | high
- [1] 性价比 | 高 Cost effectiveness | high
- [-1] 电池续航能力|不足 Battery life | Insufficient
- [1] 电池续航能力|出众 Battery life | Outstanding
- [-1] 电池续航能力|差 Battery life | Bad

排序学习中的自动标注



沃森的成功

如何获得数据？

- 与企业合作
- 组织起来

企业合作

- 发挥各自的长处
- 双赢

为什么要组织起来？

庞大的网络世界

SaaS
Yelp
SauceLabs
Salesforce.com
PaaS
IaaS
Facebook
Hardware
Google AppEngine
Google Apps

云爬

总结

- 海量数据威力巨大
- 可以做一些事情
- 组织起来
- 与企业合作