

数据挖掘 之 入课

主讲：肖 毅




EMAIL: yxiao@mail.ccnu.edu.cn

手 机：15002725359

Q Q：1061427438

个人简介



- 中国科学院，数学与系统科学研究院，预测科学研究中心，博士后 
- 香港城市大学，航贸金融研究中心，Research Fellow 
- 担任多个SCI期刊和《系统工程理论与实践》审稿人及编审
- 主持国家社科基金、教育部人文社科基金等国家和省部级项目7项
- 多项研究成果达到国际顶尖水平，发表SSCI/SCI论文十余篇 

研究方向

- (经济)信息预测、数据挖掘、商务智能

个人主页

- <http://imd.ccnu.edu.cn>



中科院预测科学研究中心





汪寿阳院士



- 第三世界科学院 (The Third World Academy of Sciences) 院士
- 国际系统与控制科学院 (International Academy of Systems and Cybernetics) 院士
- 国内外30余所知名大学的兼职教授或名誉教授
- 学术专著25部, 其中英文专著12部, SCI和SSCI论文230余篇, 引用近1700篇次
- 中央和国家有关部委提交研究报告和政策建议150余篇 (60余篇得到国务院总理等国家领导人的批示与好评), 部分政策建议被政府决策部门采纳, 成为了政府的政策
- 获中科院自然科学一等奖、教育部科技进步奖一等奖、北京市科学技术奖一等奖、省部科技进步奖/自然科学奖一等奖4次、二等奖6次、三等奖3次, 4次获Green Group Award和International Society of Multiple Criteria Decision Making Chairmanship Award
- 获得长江学者、中国科学院青年科学家奖、中国青年科技奖、国家杰出青年基金、“百千万人才计划”第一、二层次、中国科学院“百人计划”、复旦管理学杰出贡献奖
- 顶级期刊Energy Economics和Information and Management在内的15种国际期刊及8种国内重要期刊的主编、执行主编、副主编或编委



香港城市大學



- 1984年建校，2014年轻大学（<50 年）世界排名第五。



香港城市大学



- 国际化

- 语言：粤、英、中
- 师、生源：国外、香港、台湾、内地



香港城市大学-工作间



香港城市大学-工作间



全球Top20集装箱港口预测报告



2012 排名	2011 排名	港口	所属国家	集装箱吞吐量 (万 TEU)	同比 (%)
1	1	上海港	中国	2320~2360	4.9~6.2
2	2	新加坡港	新加坡	3200~3240	6.9~8.2
3	3	香港港	中国	2500~2540	2.5~4.2
4	4	深圳港	中国	2330~2370	3.3~5.1
5	5	釜山港	韩国	1755~1780	8.5~10
6	6	宁波-舟山港	中国	1660~1690	13.4~15.4
7	7	广州港	中国	1515~1540	7.3~9.1
8	8	青岛港	中国	1445~1470	10.9~12.9
9	9	迪拜港	阿联酋	1225~1250	2.1~4.2
10	11	天津港	中国	1215~1235	4.9~6.7
11	10	鹿特丹港	荷兰	1100~1210	0.0~1.7
12	13	巴生港	马来西亚	1050~1070	9.2~11.4
13	12	高雄港	中国	985~1000	2.2~3.8
14	15	汉堡港	德国	900~920	4.7~7.0
15	14	安特卫普港	比利时	875~885	1.3~2.4
16	16	洛杉矶港	美国	825~840	3.0~5.8
17	19	大连港	中国	790~800	23.5~25.1
18	17	丹戎珀拉港	马来西亚	780~790	3.4~4.8
19	18	厦门港	中国	700~710	8.3~9.9
20	21	不来梅哈芬港	德国	650~665	9.8~12.3

Forecasting for container throughput of Hong Kong Port

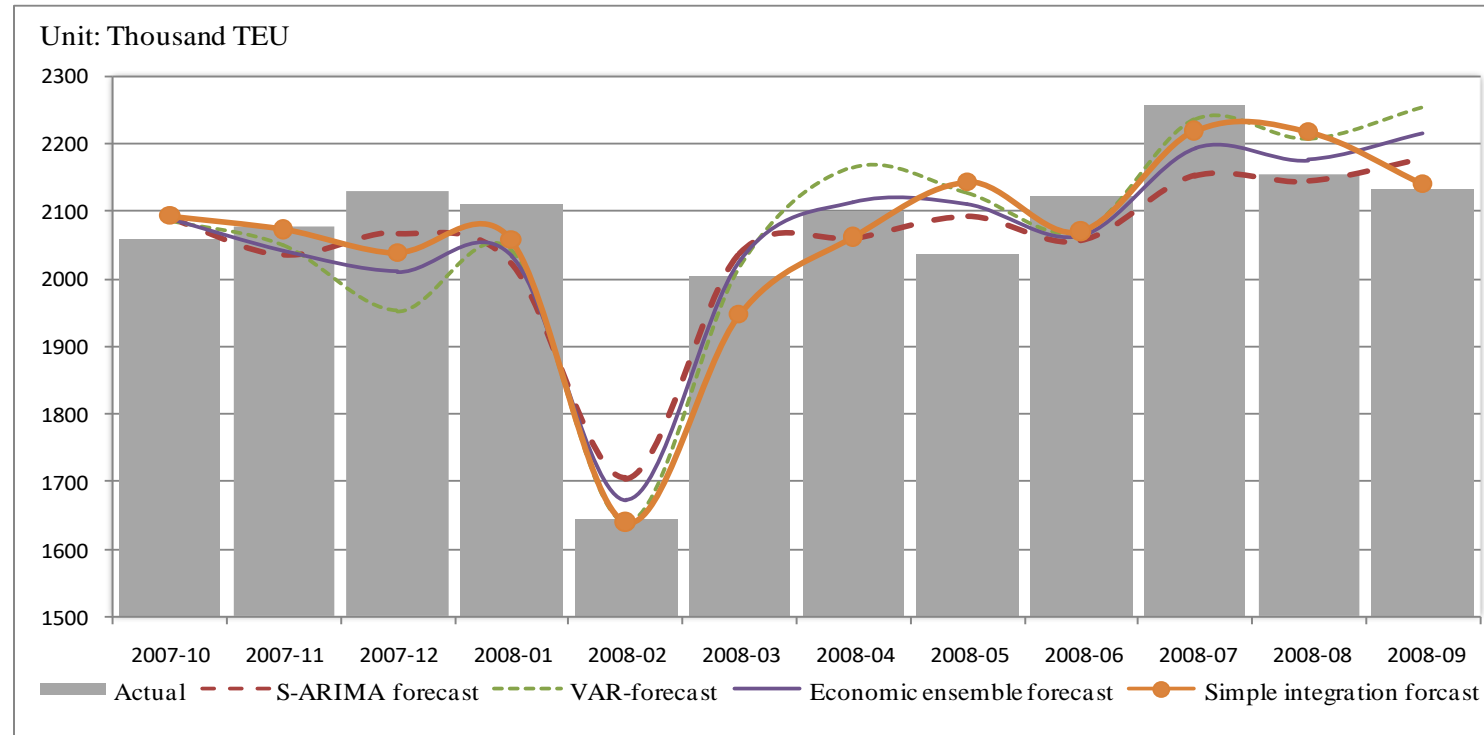


Fig. The forecasting performance of Hong Kong port container throughput

Forecasting for EUR/USD

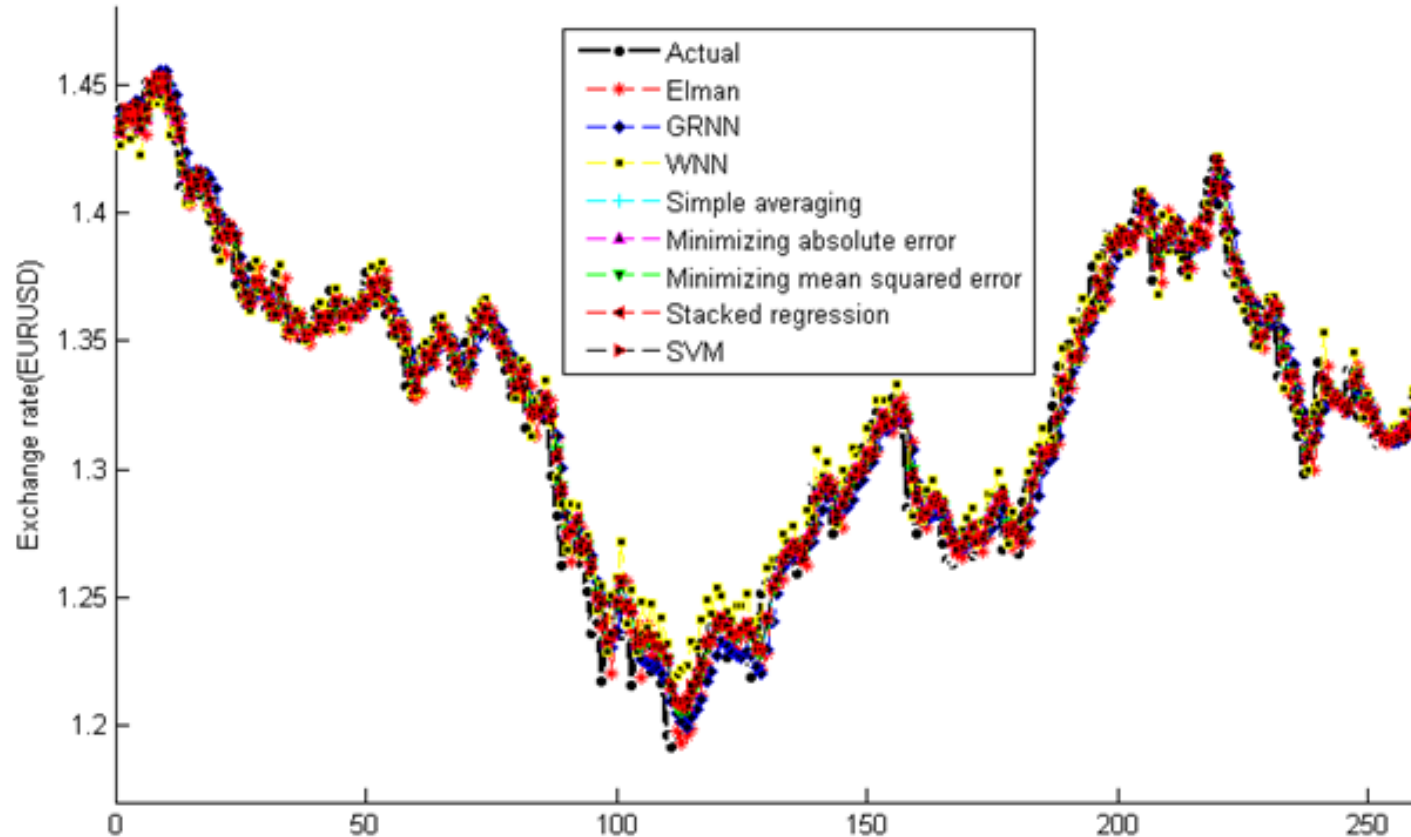


Fig. 3. Prediction of the daily exchange rates data of EUR/USD using three single base models (i.e. Elman network, GRNN and WNN), four linear ensemble methods (i.e. simple averaging, minimizing absolute error, minimizing mean squared error and stacked regression) and the proposed non-linear ensemble model in the period 1 January 2010 to 31 December 2010.

Forecasting for USD/CAD

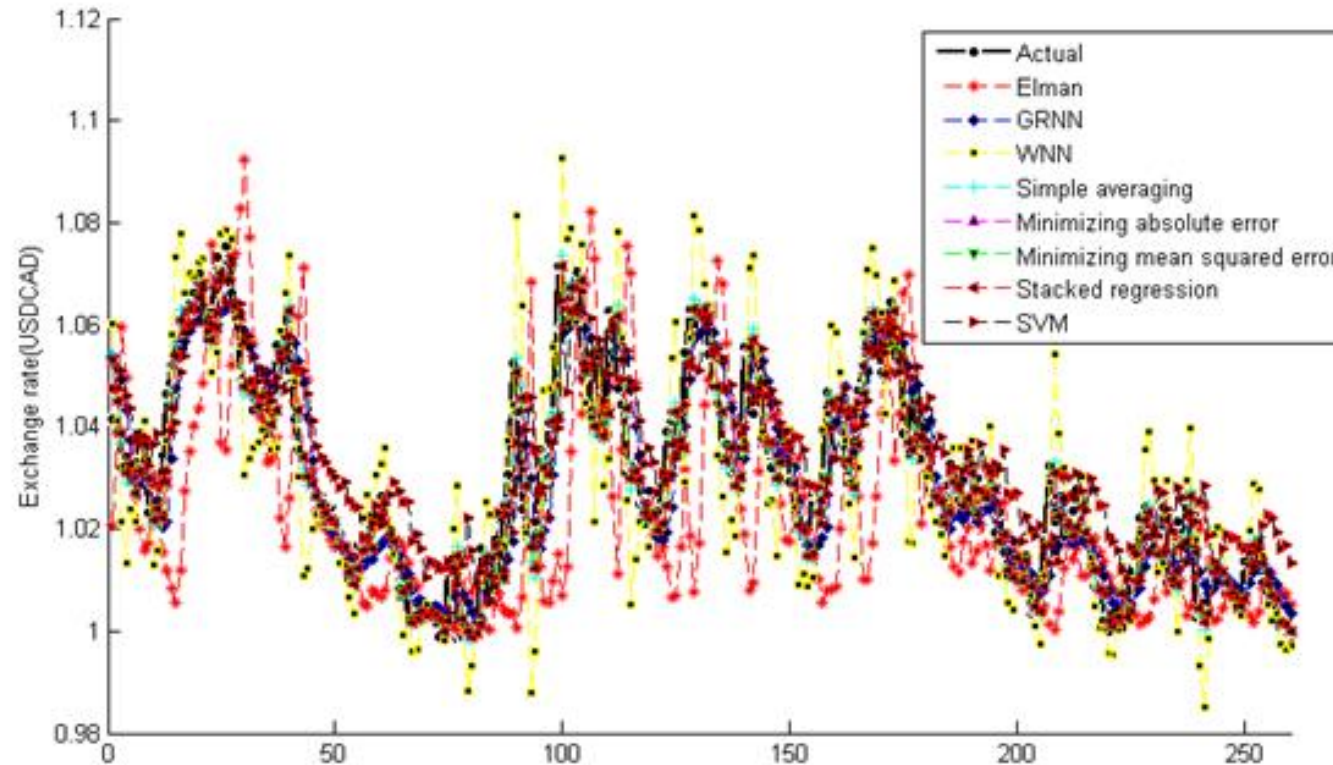
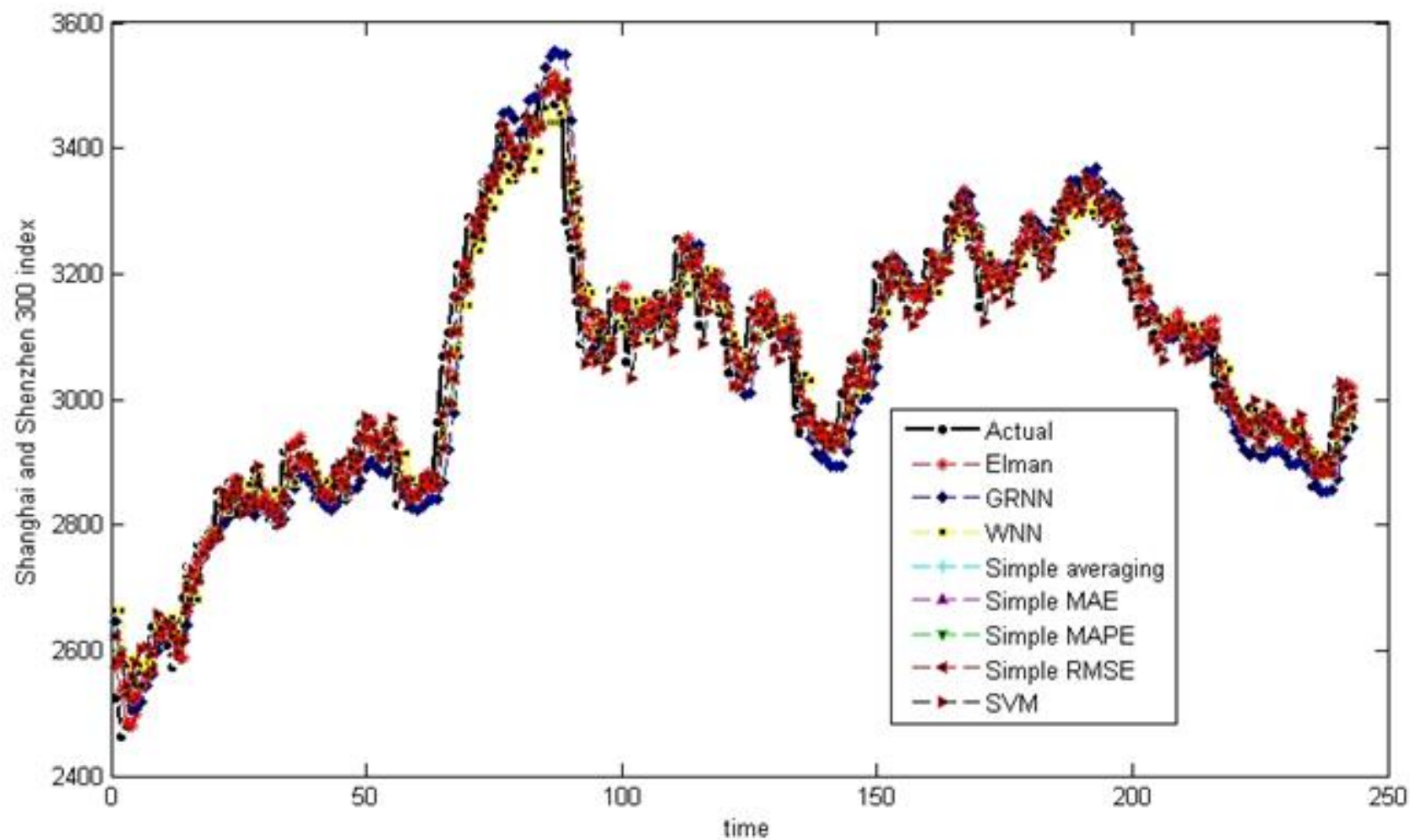
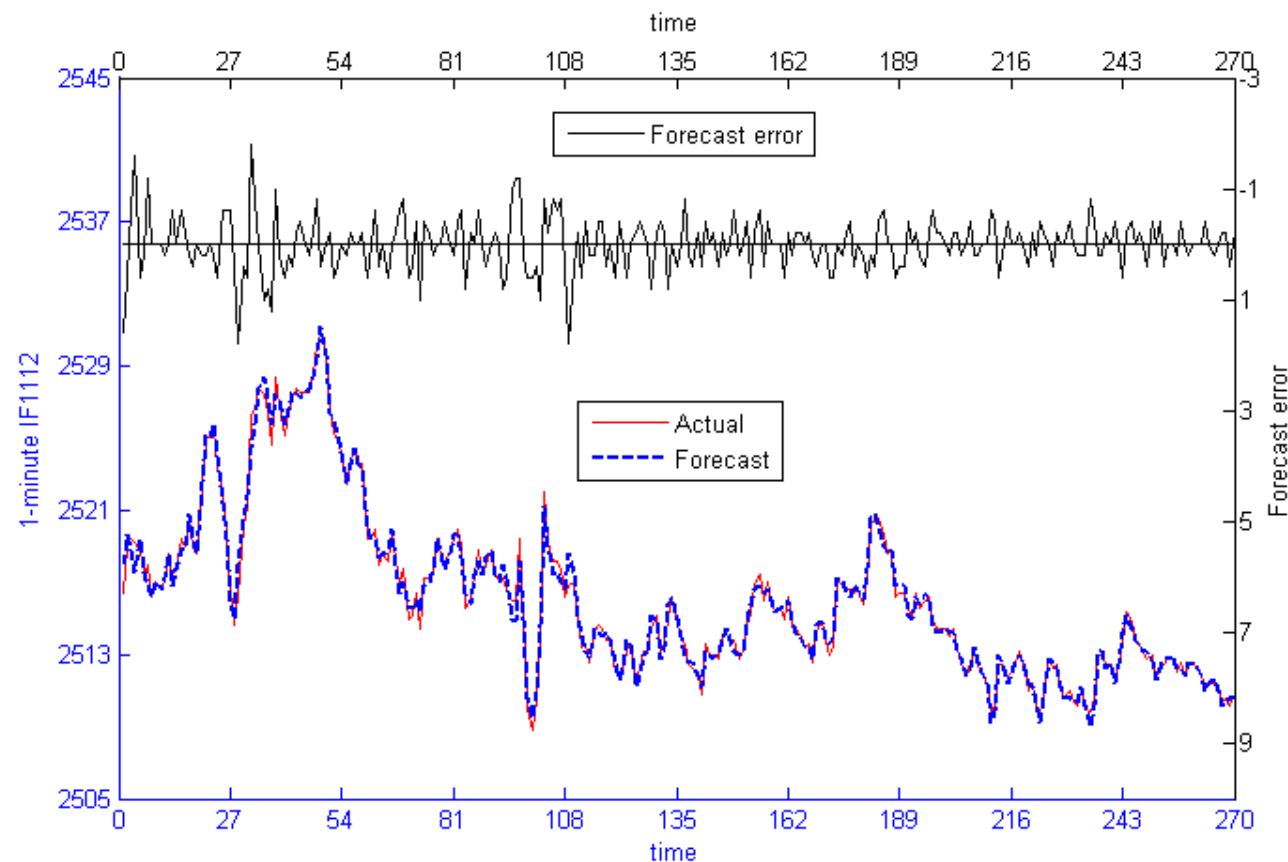


Fig. 5. Prediction of the daily exchange rates data of USD/CAD using three single base models (i.e. Elman network, GRNN and WNN), four linear ensemble methods (i.e. simple averaging, minimizing absolute error, minimizing mean squared error and stacked regression) and the proposed non-linear ensemble model in the period 1 January 2010 to 31 December 2010.

Forecasting for 沪深300



Forecasting for 股指期货



The final forecast of 1-minute HS300-SIF by the multiscale ensemble model

第一作者发表的SSCI/SCI论文



- [1] Oscillations extracting for the management of passenger flows in the airport of Hong Kong. *Transportmetrica A: Transport Science*, 2015.
- [2] Throughput estimating based port development and management policies analysis. *Maritime Policy & Management*, 2015.
- [3] Application of multiscale analysis-based intelligent ensemble modeling on airport traffic forecast. *Transportation Letters: the International Journal of Transportation Research*, 2015, 7(2), 73-79.
- [4] Multiple dimensioned mining of financial fluctuation through radial basis function networks. *Neural Computing and Applications*, 2015, 26(2), 363-371.
- [5] Feature-selection-based dynamic transfer ensemble model for customer churn prediction. *Knowledge and Information Systems*, 2015, 43, 29-51.
- [6] Domestic air passenger traffic and economic growth in China: Evidence from heterogeneous panel models. *Journal of Air Transport Management*, 2015, 42, 95-100.
- [7] A neuro-fuzzy combination model based on singular spectrum analysis for air transport demand forecasting. *Journal of Air Transport Management*, 39, 1-11, 2014.
- [8] A multiscale modeling approach incorporating ARIMA and ANNS for financial market volatility forecasting. *Journal of Systems Science and Complexity*, 27(1), 225-236, 2014.
- [9] A transfer forecasting model for container throughput guided by discrete PSO. *Journal of Systems Science and Complexity*, 27(1), 181-192, 2014.
- [10] Ensemble ANNs-PSO-GA approach for day-ahead stock E-exchange prices forecasting. *International Journal of Computational Intelligence Systems*, 2014, 7(2), 272-290.



为什么需要数据挖掘



- 商业观点

- 数据来源：网页数据，电子商务交易记录，在商店的购物统计，银行/信用卡
- 提供更个性化的服务以取得优势（例如：在客户关系管理方面）

- 科学观点

- 数据在以非常高的速度进行采集和储存(GB/小时)
- 数据挖掘或许可以帮助科学家
 - 在数据分类和数据细分方面
 - 在假说的形成方面



“人类正被数据淹没，却饥渴于信息” ——John Naisbitt（未来学家）

为什么需要数据挖掘



- 数据里经常有一些并不是很明显的“隐藏”的信息
- 找到这些“隐藏”的信息非常困难

“人类正被数据淹没，却饥渴于信息” ——John Naisbitt (未来学家)

- “需要是发明之母” ——数据挖掘——大量数据集的自动分析

数据挖掘是什么



- 数据挖掘——Data Mining (DM)
- 数据挖掘：从大量数据中发现潜在的有用模式（信息、知识、规律、模型）的过程
 - 猎人在动物迁徙的行为中寻找模式
 - 农夫在庄稼的生长中寻找模式
 - 政客在选民的意见中寻找模式
 - 医生在病人的症状中寻找模式

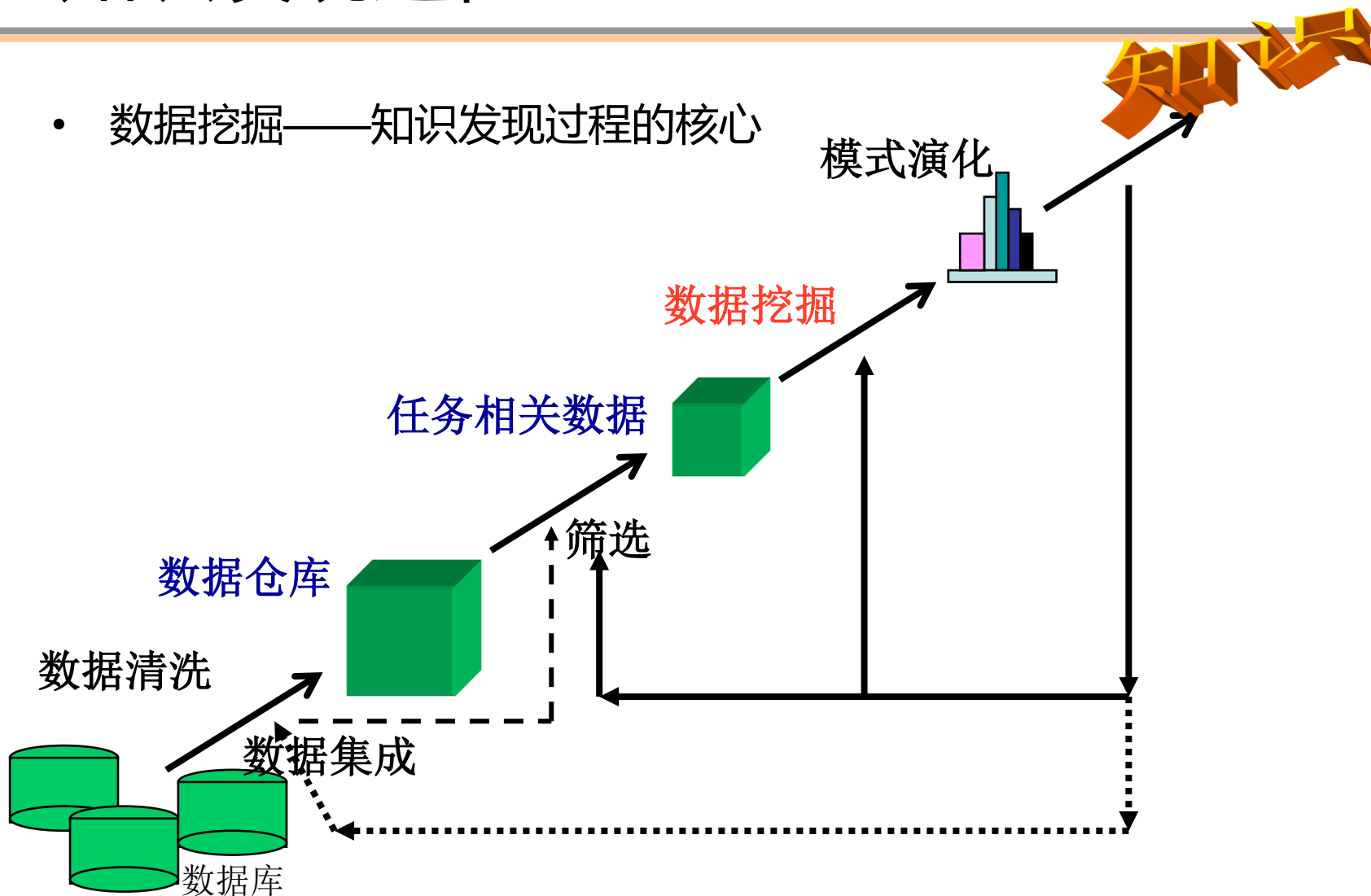


“人类正被数据淹没，却饥渴于信息” ——John Naisbitt (未来学家)

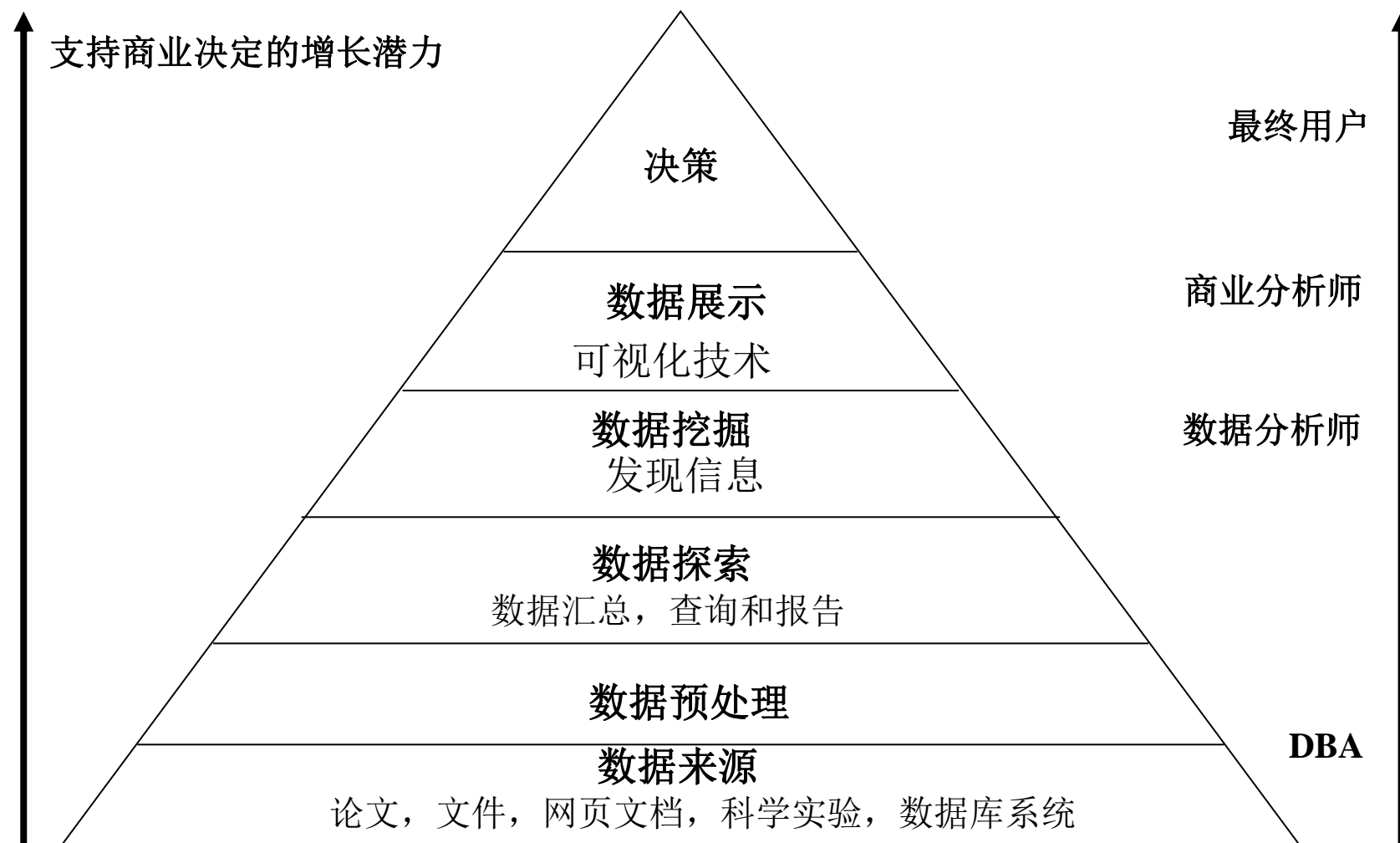


知识发现过程

- 数据挖掘——知识发现过程的核心



数据挖掘和商业智能





数据挖掘：多学科的合流

