智能金融大数据投研平台

林健武

清华大学深圳国际研究生院 · 教 授

清華大學 深圳国际研究生院
Tsinghua Shenzhen International Graduate School

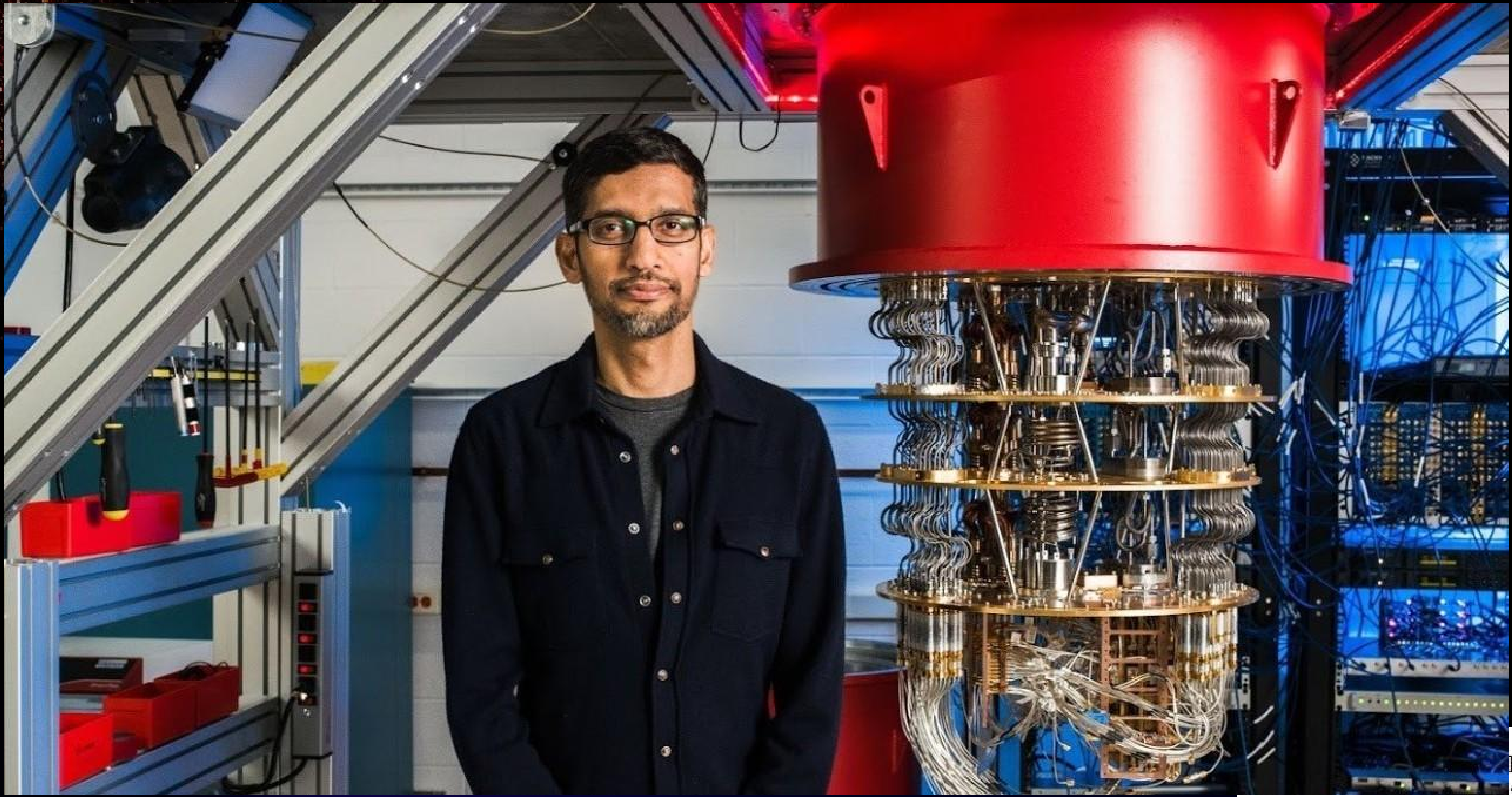
金融科技的发展历程


决策:
大数据金融
投资:众筹
融资:P2P网贷
支付:第三方支付
流通:信息化金融机构
货币:数字货币
金融科技3.0
金融科技2.0
金融科技1.0

金融科技3.0框架


金融科技3.0 ABBC
计算资源
云计算
征信
大数据
资金
记录跟踪
区块链
决策
人工智能
支付
借贷
投资
OPEN
贷
股
商品、服务
债权
有抵押、无抵押、自融
股权
一级市场
二级市场

谷歌量子计算突破登Nature封面，据说200秒顶超算10000年



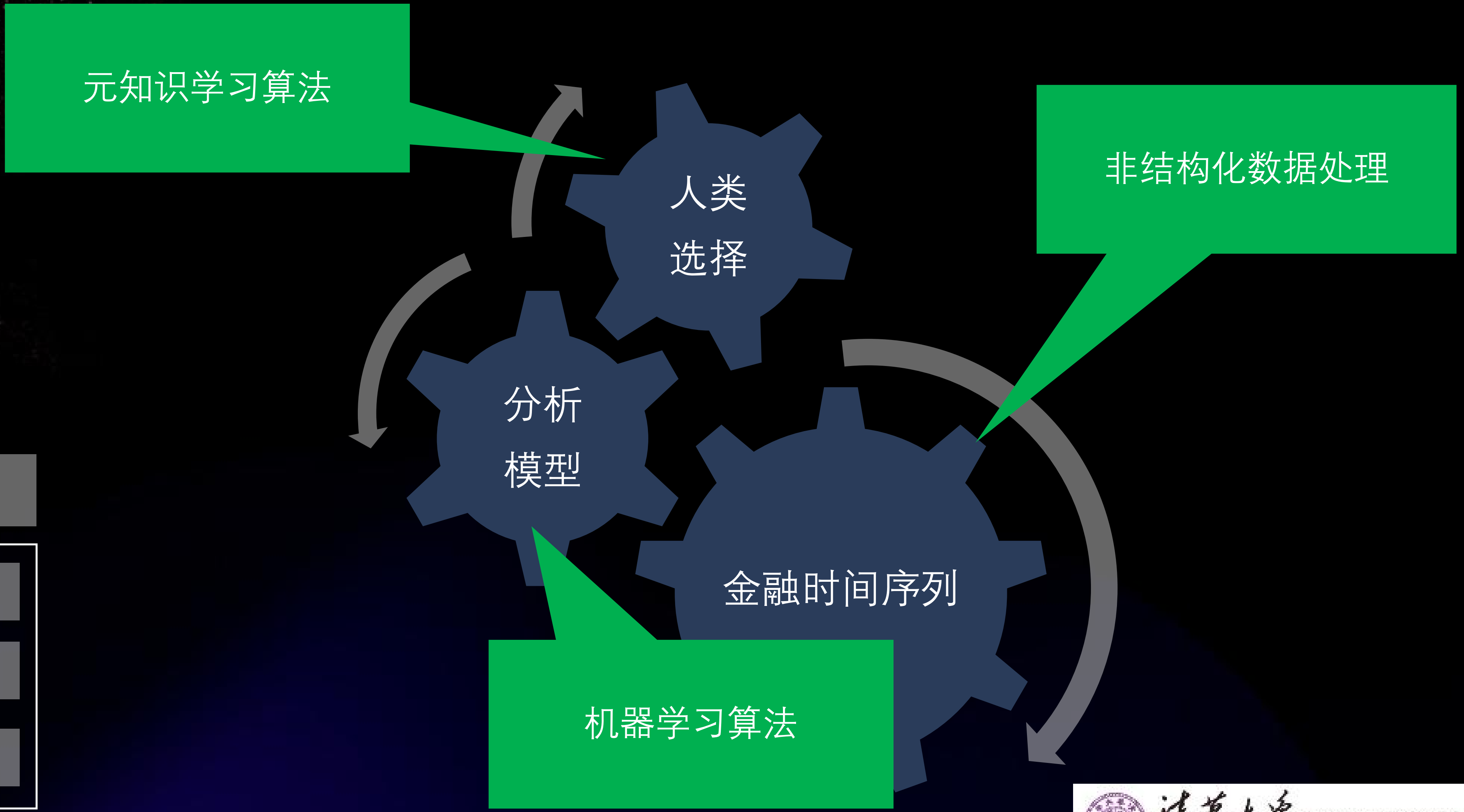
目录

CONTENTS

清華大學 深圳国际研究生院
Tsinghua Shenzhen International Graduate School

智能金融大数据投研平台框架

智能金融大数据投研平台框架



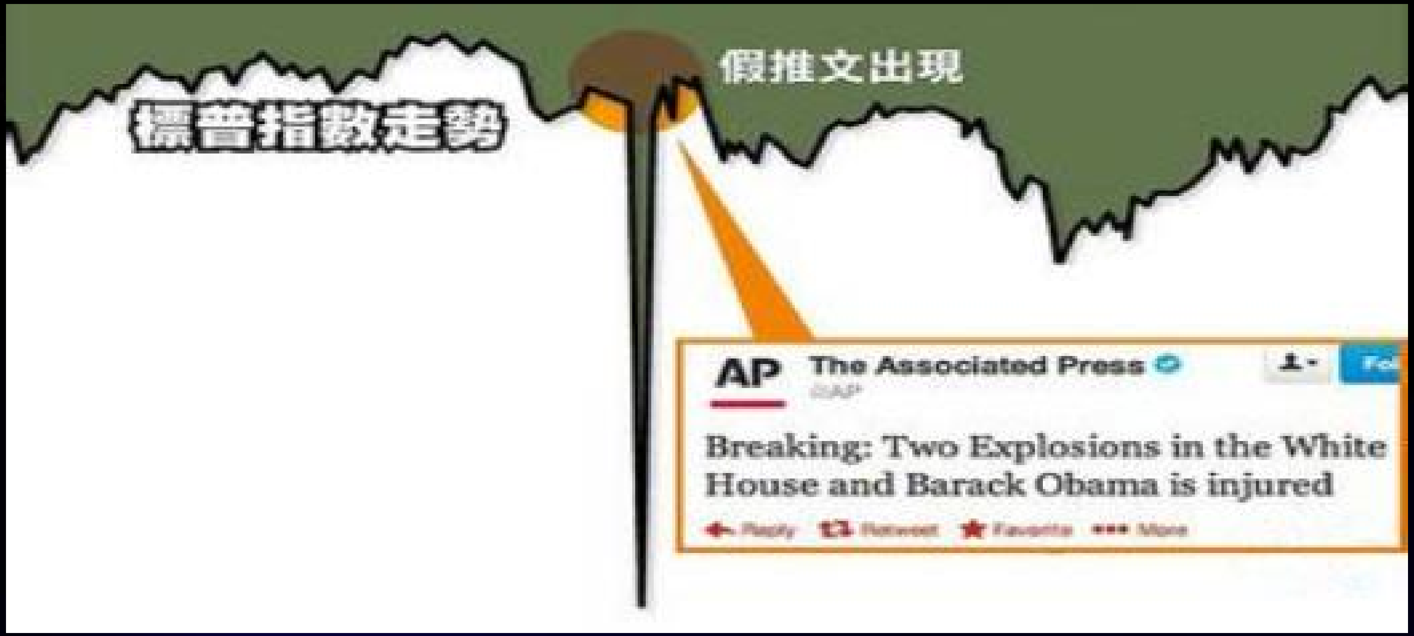
分析模型

- 输入数据
- 特征提取
- 算法结构

非结构化数据处理

新闻非结构化数据的实例

- 2013年4月23日，美联社发布推文称“两枚炸弹在白宫爆炸，奥巴马总统被炸伤”
- 标准普尔500指数暴跌1%，一分钟内1360亿美元瞬时蒸发。



金融文本情感分析

Applications:

Mainly in E-commerce

Sentiment Analysis

Approaches
- Lexicon-based,
- Regular Machine Learning,
- Deep Learning,
- Hybrid

3 levels:
- Document level,
- Sentence level,
- Aspect level

Financial Analysis

Current main approaches
- Fundamental Analysis
- Technical Analysis

Market sentiment: remarkable influences on price trends, trading volumes, volatility and potential risks

Application of FSA: Thomson Reuters News Analytics(TRNA) scores

各种金融文本的对比

In FSA, the sources' influences are important and the characteristics should be considered.

News cover more kinds of events than corporate disclosure.

Social media are widely analyzed despite its noise.

Some researches tend to link the sentiment of news and micro-blogs and improve the conveyance of news-contained sentiment on micro-blogs.

Different financial news media have their own characteristics.

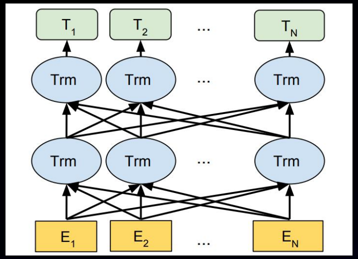
The financial articles' origin had different influences on investors.

In social media, user-profile features should be considered.

情感分析算法

Word embedding

- Word2Vec
- Global Vector（GloVe）： contain word order and global information
- Contextualized word embedding
  e.g.ELMo， GPT
  Milestone：BERT (Bidirectional Encoder Representation from Transformers)



Trend：GPT-2 use a large and diverse dataset and a very deep neural network

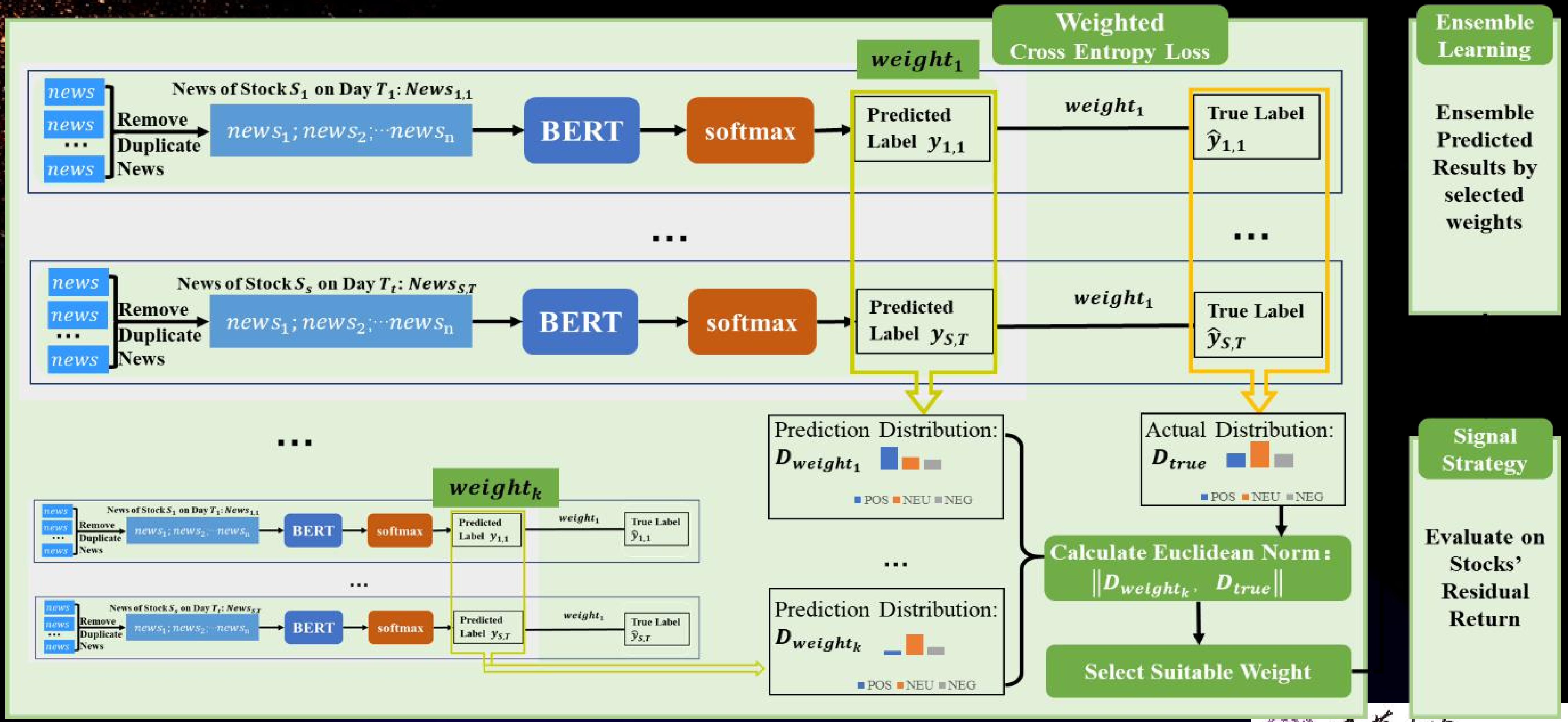
CNN,RNN,LSTM,Bi-LSTM

- Convolution Neural Network (CNN) can effectively capture local correlations of spatial or temporal structures.
- RNN can handle the long dependency of sequential better than CNN.
- Both CNN and Bi-LSTM[66] can be combined to learn the sequential correlations and extract features in a parallel way.

Attention Mechanism

- The attention mechanism allows the model to focus on the needed parts.
- Transformer is designed solely on attention mechanisms.
- Good at important information recognition from both sentence and aspect, position awareness and modeling the relationships between aspect terms.
- Researchers introduced the attention mechanism to explore the correlations between a financial aspect and the context.
- How to define a broadly accepted aspect in FSA needs to be discovered.

基于BERT的金融文本数据特异质收益分析模型



清华大学 深圳国际研究生院
Tsinghua Shenzhen International Graduate School

Enhancement Learning

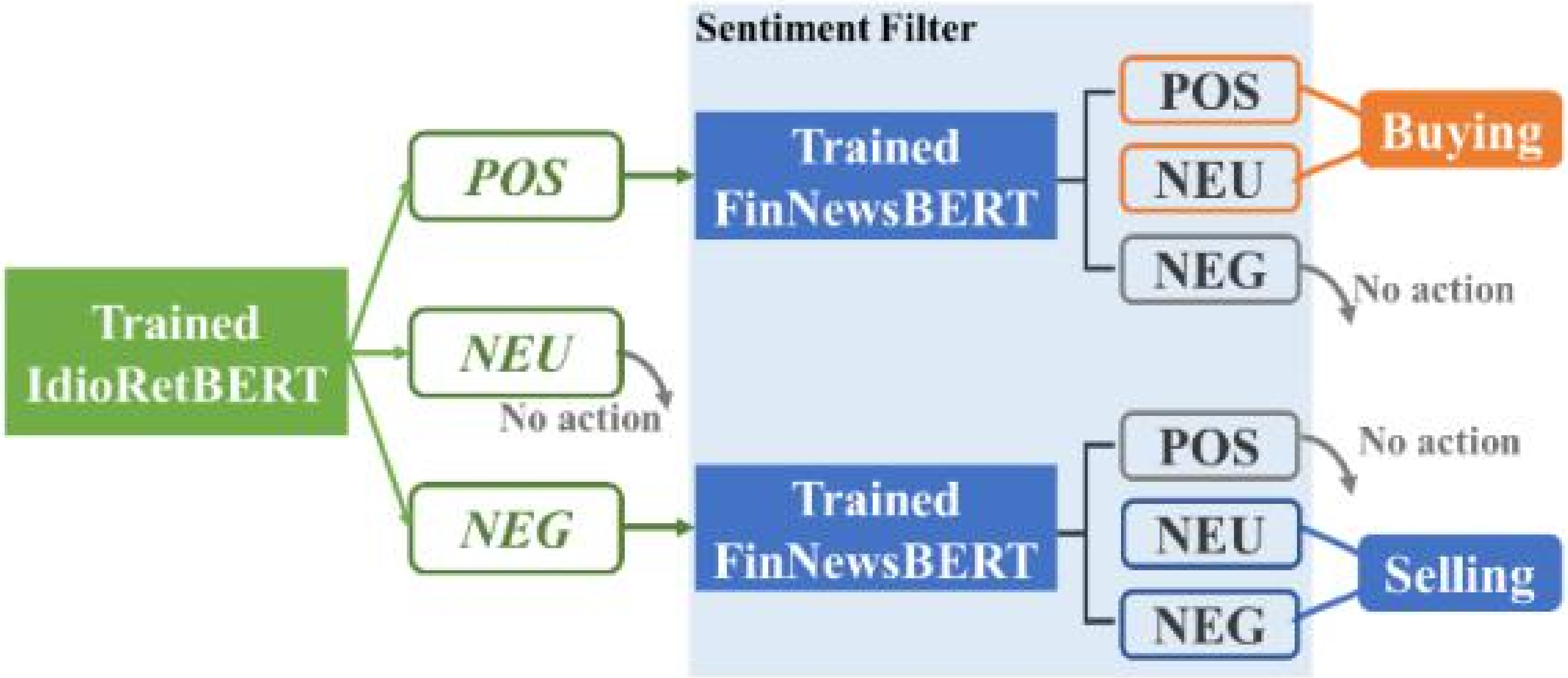
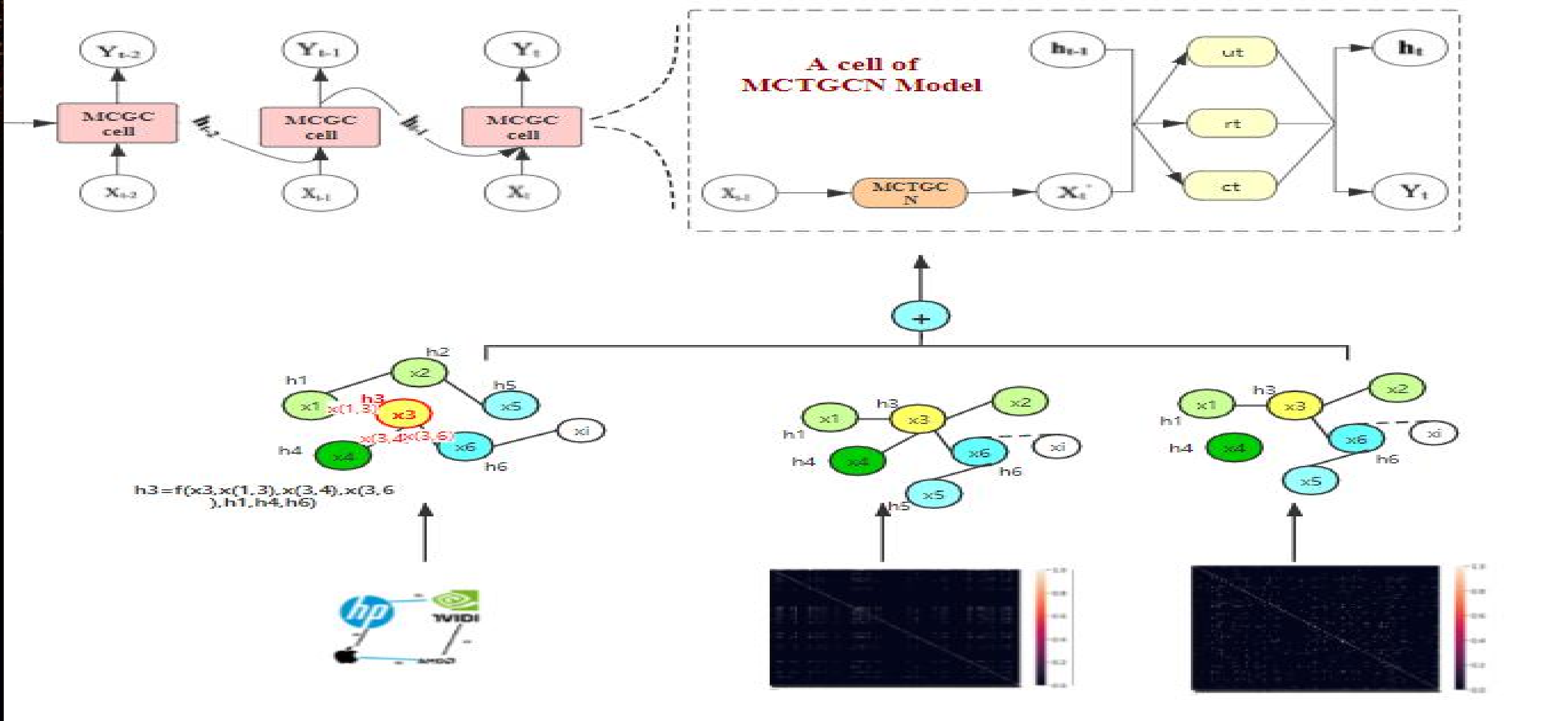


Fig. 9. The task-oriented perception-enhanced approach of EL

图相关金融文本数据分析模型



专利大数据平台

专利是科技公司的护城河、防火墙，也是科技公司创新能力的集中体现。一直以来，学术研究焦点基本在于专利数量、专利分类以及专利引用量三个方面，经过众多学术研究表明，专利大数据是科创企业的试金石！

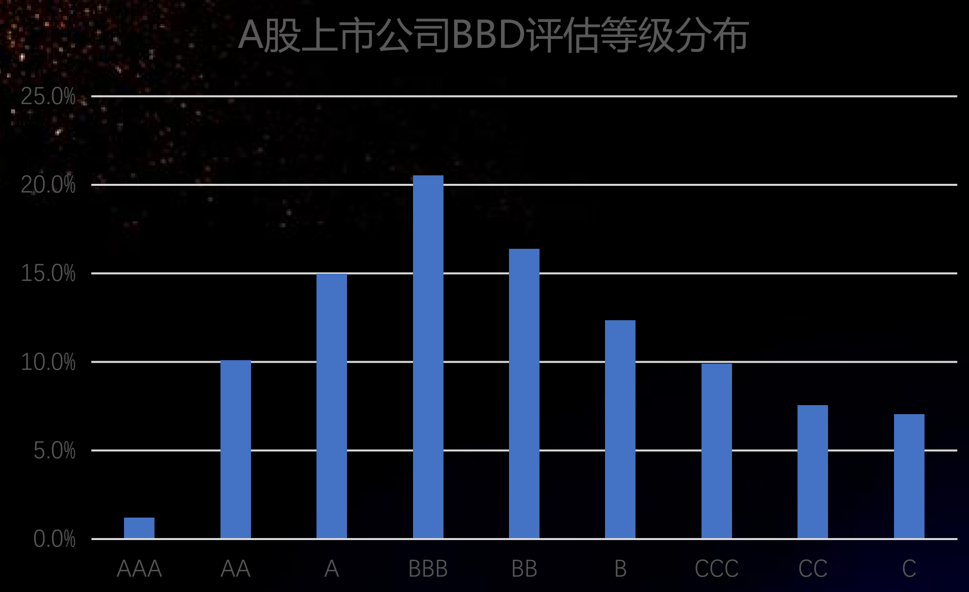
- 对基本面的影响：专利数量与公司市值呈显著正相关
- 对公司股价的影响：专利数量和专利引用量以及股票收益呈现显著正相关关系

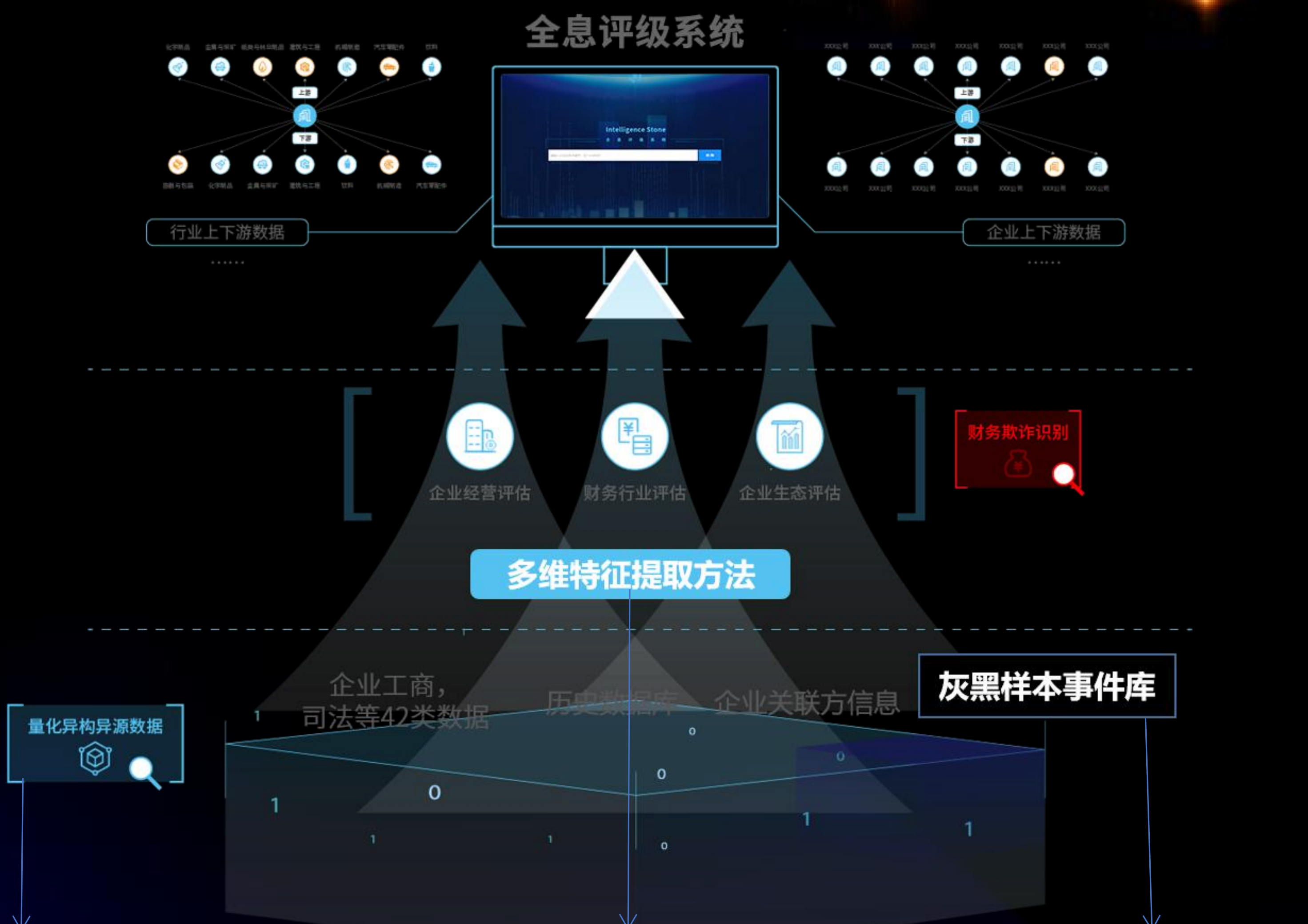
- 将专利研发成果细分为开发利用（短期）与探索研发（长期）两大类，开发利用类的专利可以给公司的业收入带来显著正面影响，而探索研发类的直接作用并不明显。

- 对基本面的影响：专利引用量的增加可以帮助正向预测公司未来的现金流和收益
- 对公司股价的影响：同一公司对内部过往专利的引用（Self-atation）比对公司外部其他专利的引用对公司市值影响更大

信用大数据平台

- 等级区分度明显
- 评估等级为AA以上的上市公司仅占上市公司的10%左右（300至400家），等级为AAA的公司仅占上市公司的1%左右。





引入“生态风险评估”，将异构异源的数据融合分析，降低对财务数据和专业分析师的依赖

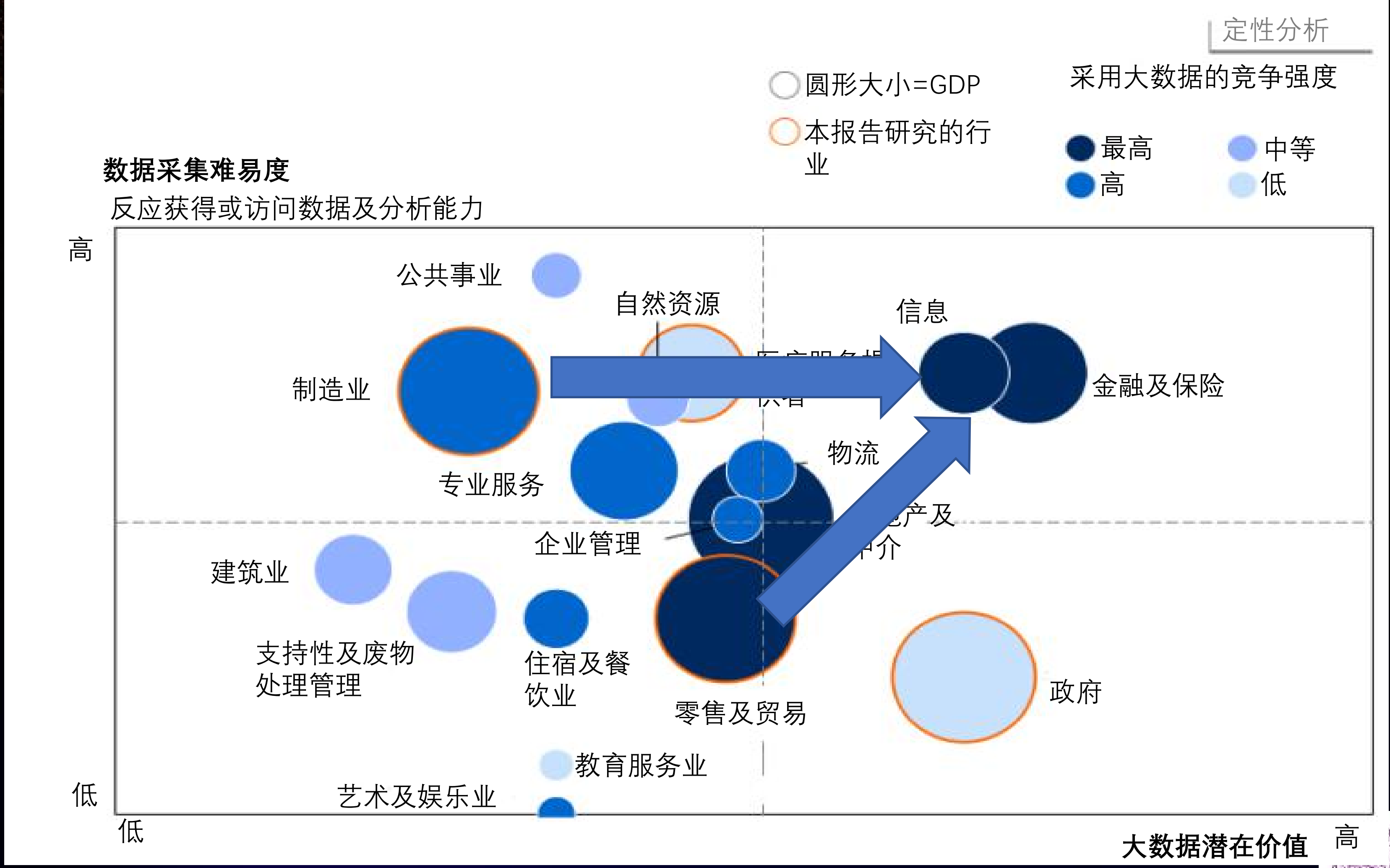
基于多维数据的融合分析，采用融合量化处理非结构化与结构化数据的特征提取方法，可以对评估标的进行有效的特征提取，实现对企业风险的多维度评估

拥有大量灰黑样本，可以解决债券市场违约样本不足的问题

清華大學 深圳国际研究生院 Tsinghua Shenzhen International Graduate School

大数据 – 潜在价值

不同行业使用大数据的能力和获得的价值有所不同



机器学习算法

数据量与建模方法

如何进行机器学习

统计学	算法结构	特性提取	输入数据
遗传算法	算法结构	特征提取	输入数据
深度学习	算法结构	特征提取	输入数据
自动机器学习	算法结构	特征提取	输入数据

人工 → 机器

清華大學 深圳国际研究生院
Tsinghua Shenzhen International Graduate School

Prior Knowledge Distillation (PKD) – High noise

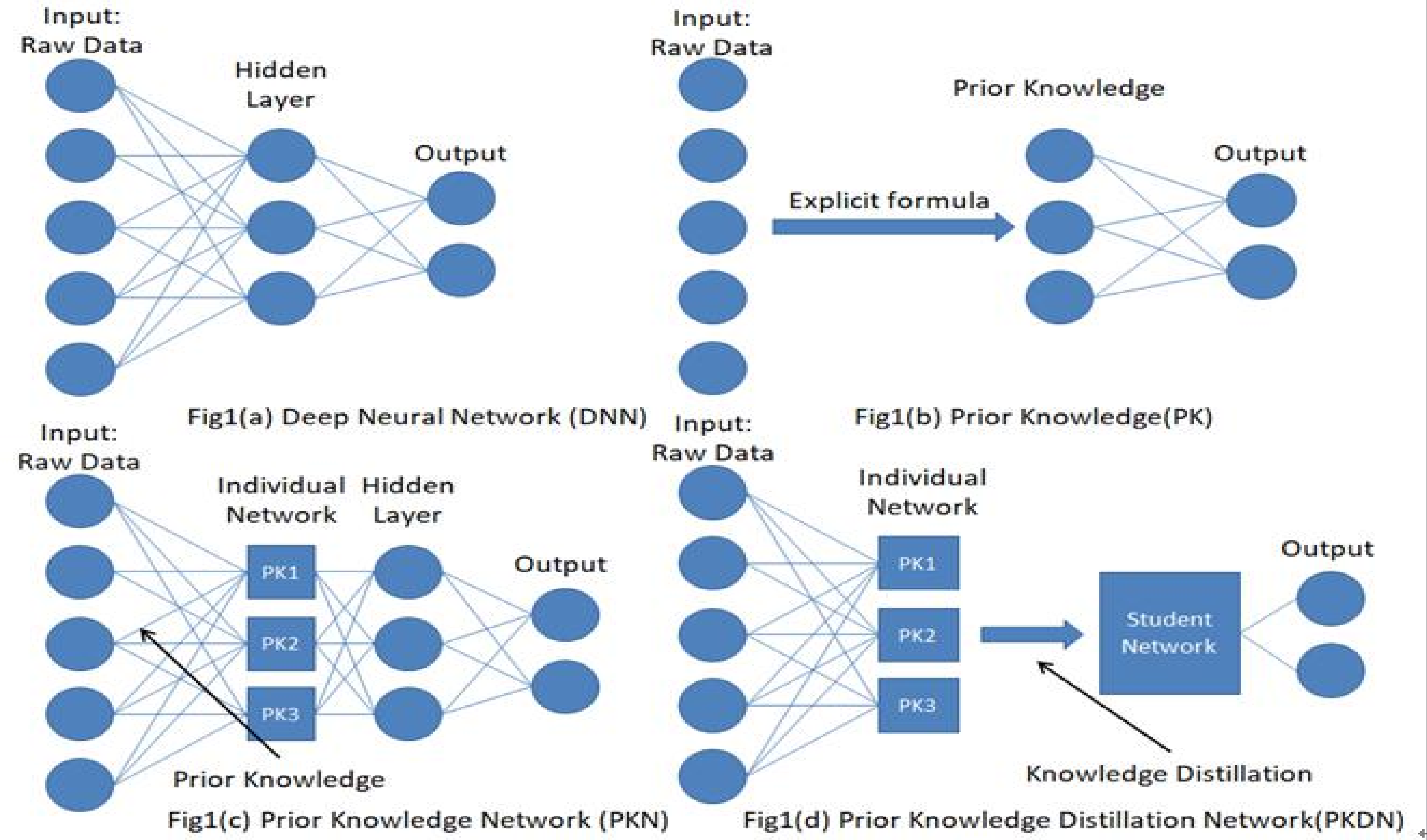


Figure1. Network structures

Neural Network-based Automatic Factor Construction

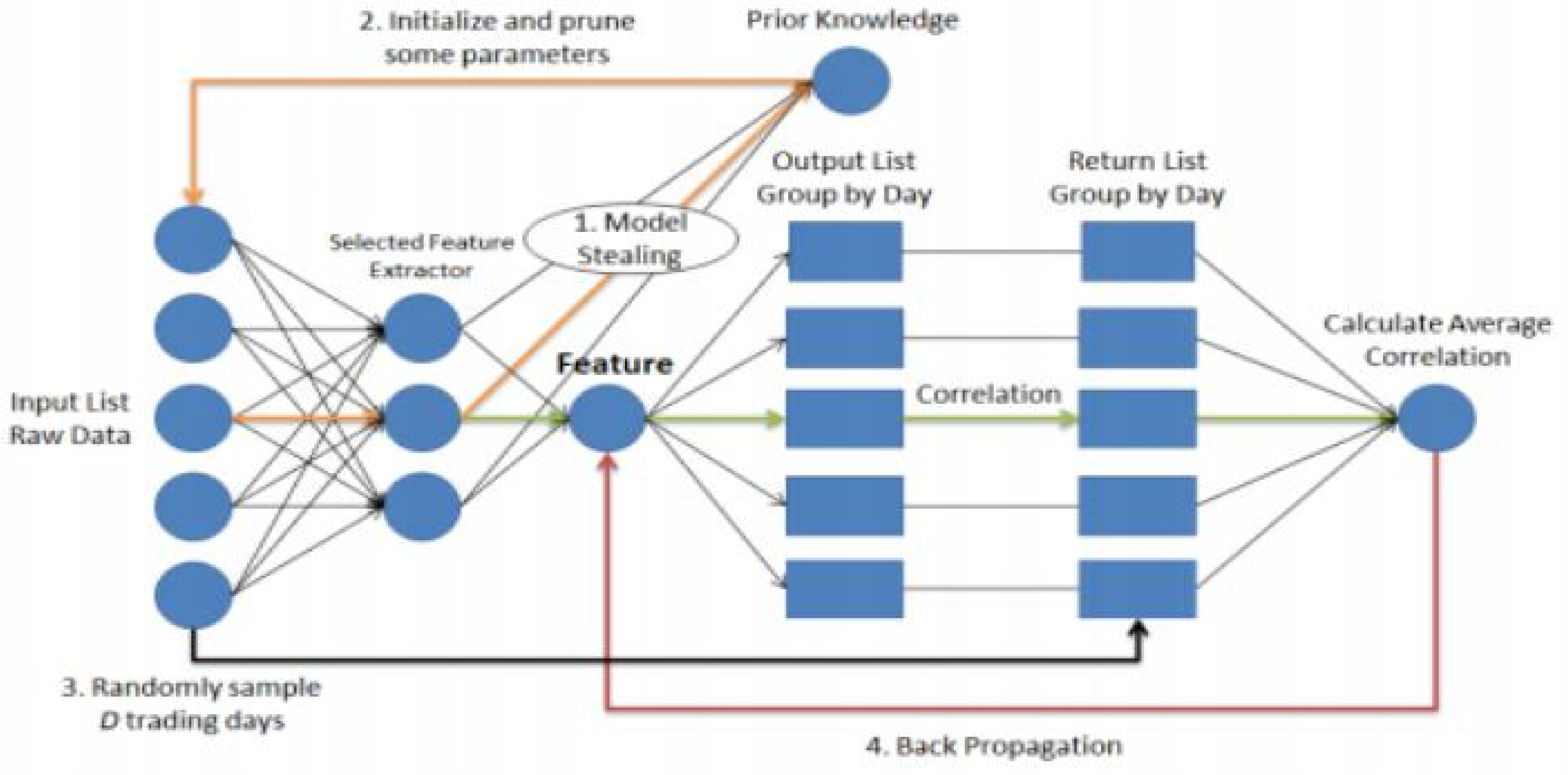
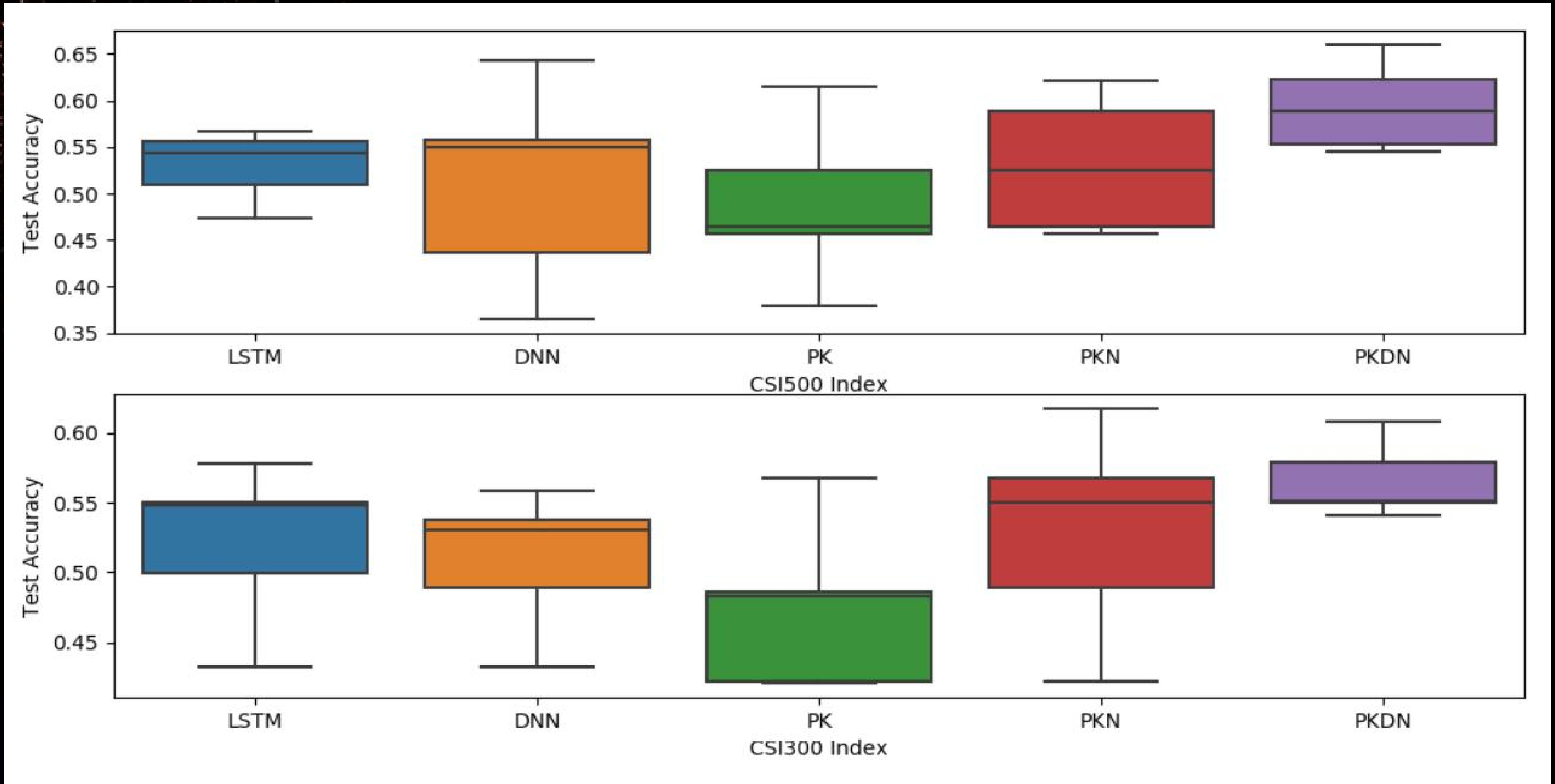


Figure 2. Neural Network-based Automatic Factor Construction's framework

Numerical tests for PKD



元知识学习算法

元知识学习计算时间

- Time: 100 strategies x 1000 training sets x 10000 parameter sets x 100 seconds per simulation = 1e11 seconds

  more than 3000 years

- Parallel computing with 100 CPUs

  more than 30 years

- Apply genetic algorithm to reduce parameter sets by 10 times

  more than 3 years

- Use powerful GPU to speed up simulation by 10 times

  more than 3 months

Asymptotic Ordinal Optimization

- **Partial sorting**

Selecting the best

- **Ordinal Optimization**

The underlying philosophy is to obtain good estimates through ordinal comparison while the value of an estimate is still very poor

- **Asymptotic Optimization**

Leveraging previous simulation performance

Asymptotic Meta Learning (AML)

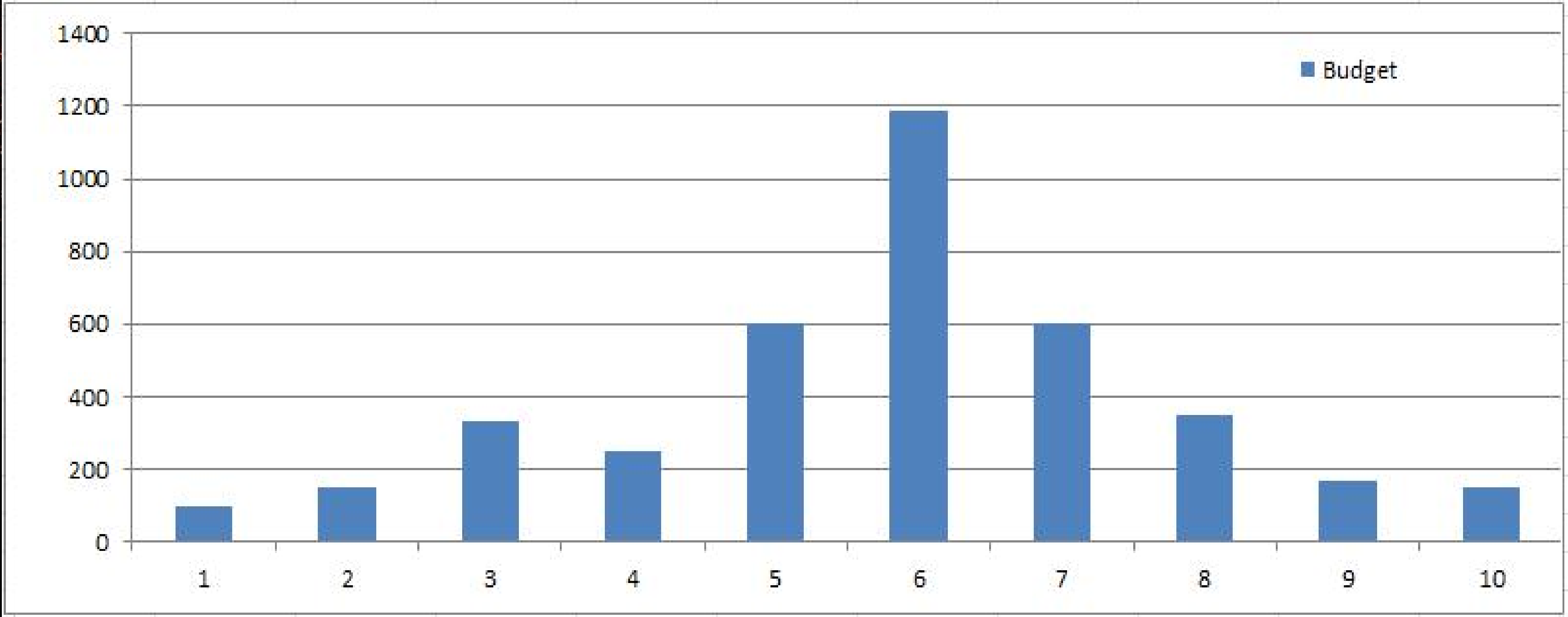
Theorem 1. *Given a total number of computing time T to be allocated to k base learners whose performance is depicted by cross validation performance $L(\theta_1, \xi)$, $L(\theta_2, \xi)$, …, $L(\theta_k, \xi)$ with means $J(\theta_1)$, $J(\theta_2)$, …, $J(\theta_k)$, and finite variances $\sigma_1^2, \sigma_2^2, \ldots, \sigma_k^2$ and average learning time per learner $C_1, C_2, \ldots C_k$, respectively, as $T \to \infty$, the Approximate Probability of Correct Selection (APCS) of Meta learning can be asymptotically maximized when*

1) $\frac{N_i}{N_j} = \left(\frac{\left(\frac{\sigma_i}{\delta_{b,i}}\right)}{\left(\frac{\sigma_j}{\delta_{b,j}}\right)}\right)^2, i, j \in \{1, 2, \ldots k\}, i \neq j \neq b$

2) $N_b = \sigma_b \sqrt{\sum_{i=1, i \neq b}^{K} \frac{C_i N_i^2}{C_b \sigma_i^2}}.$

Where N_i is the number of samples allocated to base learner i, $\delta_{b,i} = \overline{J_b} - \overline{J_i}$, and $\overline{J_b} \geq \max_i \overline{J_i}$. And we assume $N_b >> N_i$.

AML Bootstrapping



Mean	8	7	6	5	4	3	4	5	6	7
std	5	6	10	5	6	7	6	7	5	6

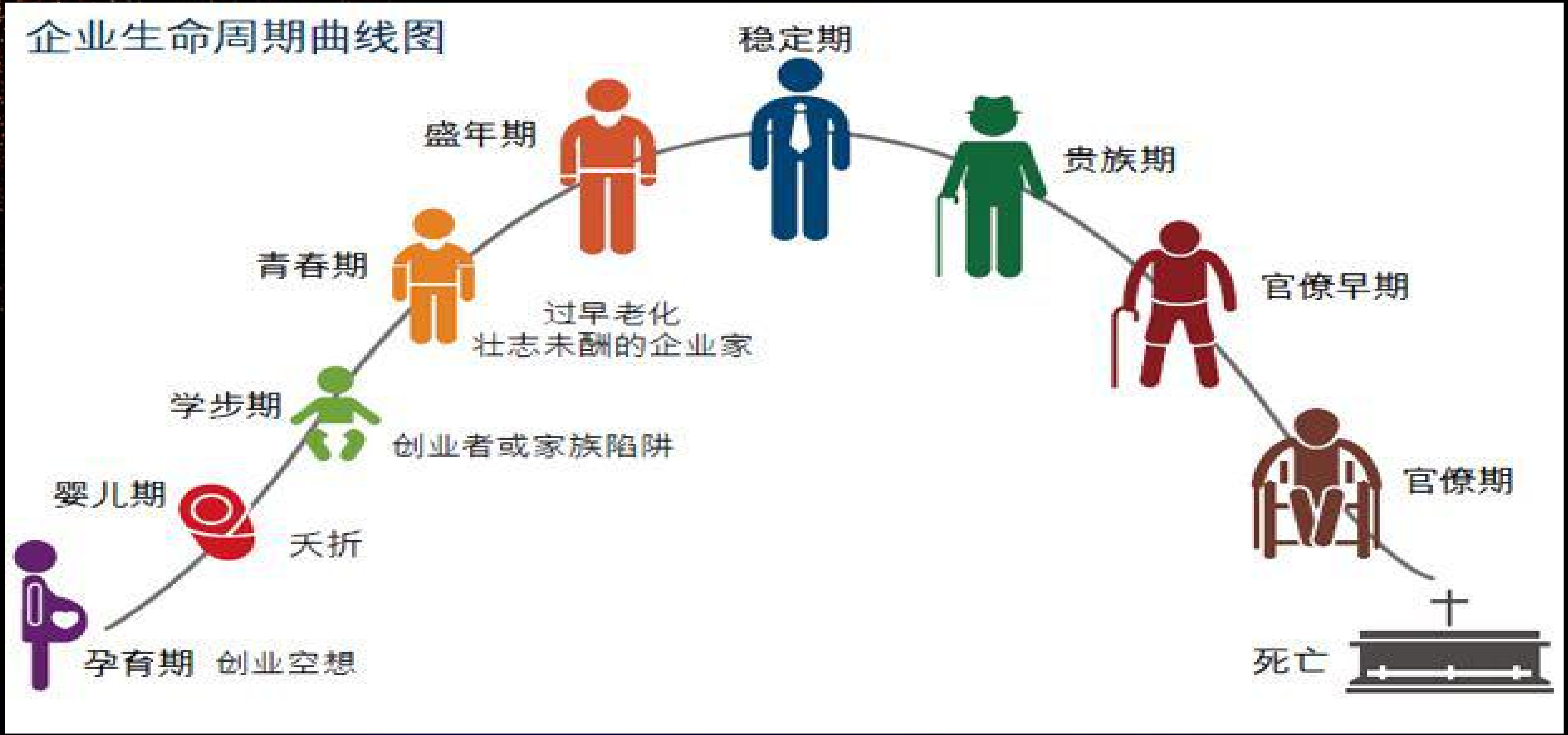
AML Numerical Experiments

TABLE 1

Results of numerical experiments for four Meta learning algorithms under different case

Validation time with 95% P{CS} Cutoff (seconds)	AML	OCBA	Equal	Greedy
Case 1	3.69	4.30	10.18	4.85
Case 2	12.51	17.03	42.00	13.93
Case 3	7.41	7.85	21.78	8.72
Case 4	10.26	12.64	31.08	29.71
Case 5	16.51	19.81	36.85	48.86
Case 6	5.65	7.49	13.26	7.31

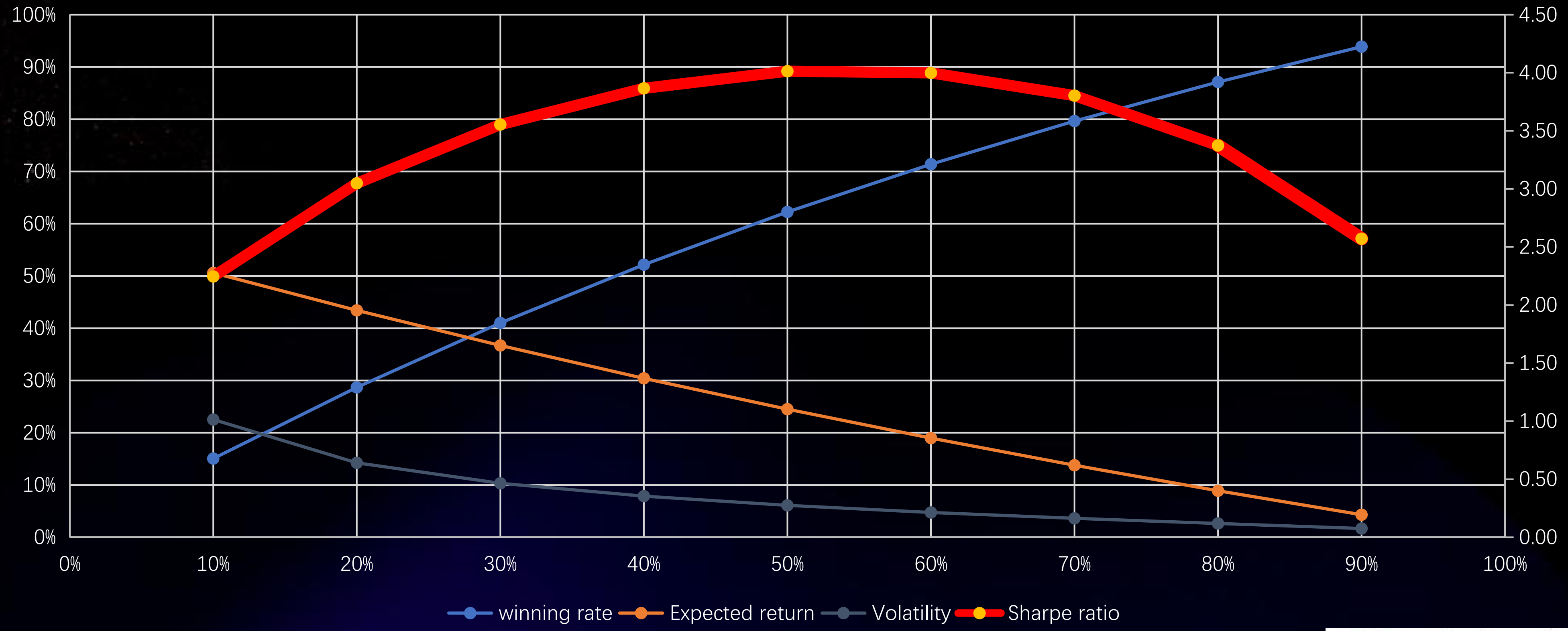
生命周期算法 – 群体博弈



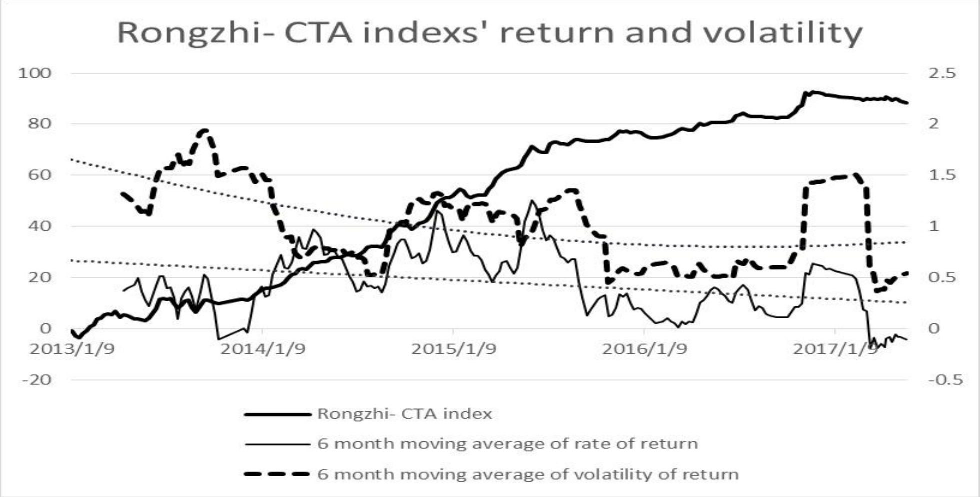
清華大學 深圳国际研究生院 Tsinghua Shenzhen International Graduate School

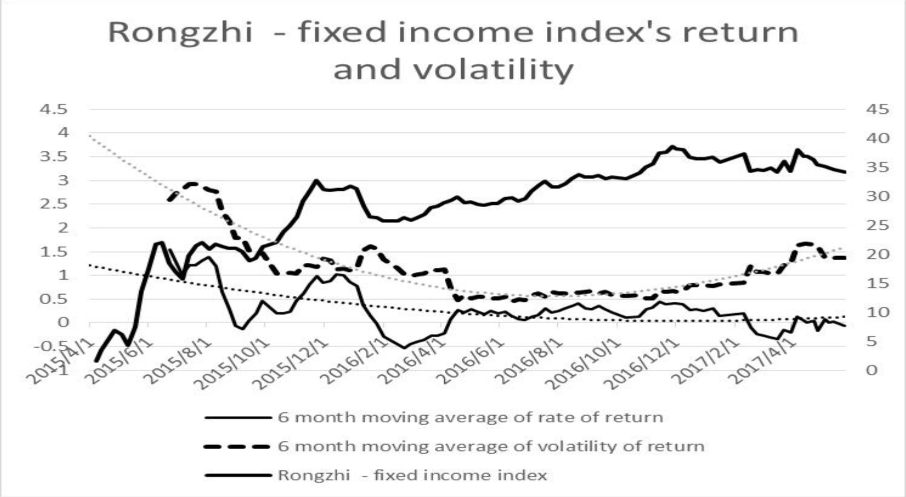
生命周期模型

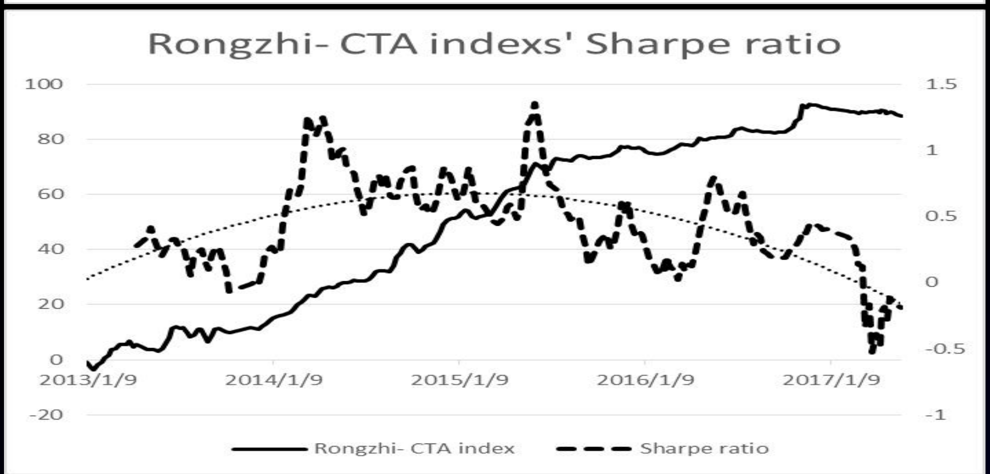
- 对于单个策略，在市场总资金和机构投资者信息系数保持不变的情况下，机构投资者的收益的夏普呈现Alpha周期的现象

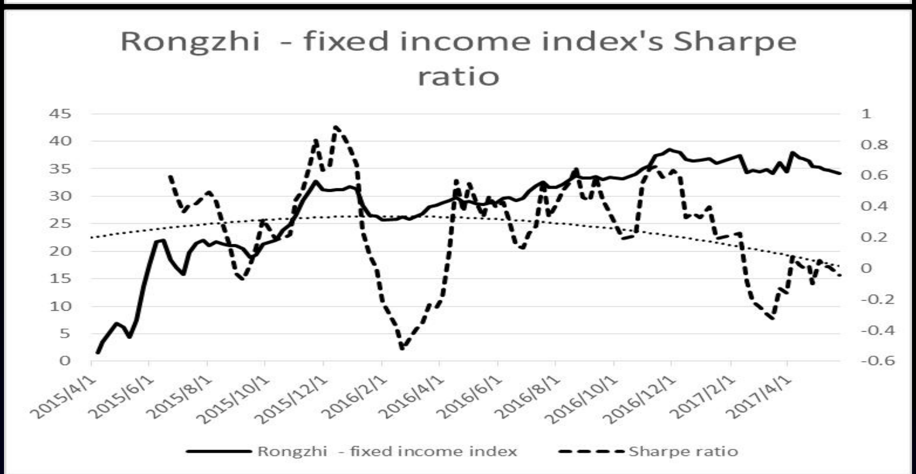


实证研究


Rongzhi- CTA indexs' return and volatility
Rongzhi- CTA index
6 month moving average of rate of return
6 month moving average of volatility of return

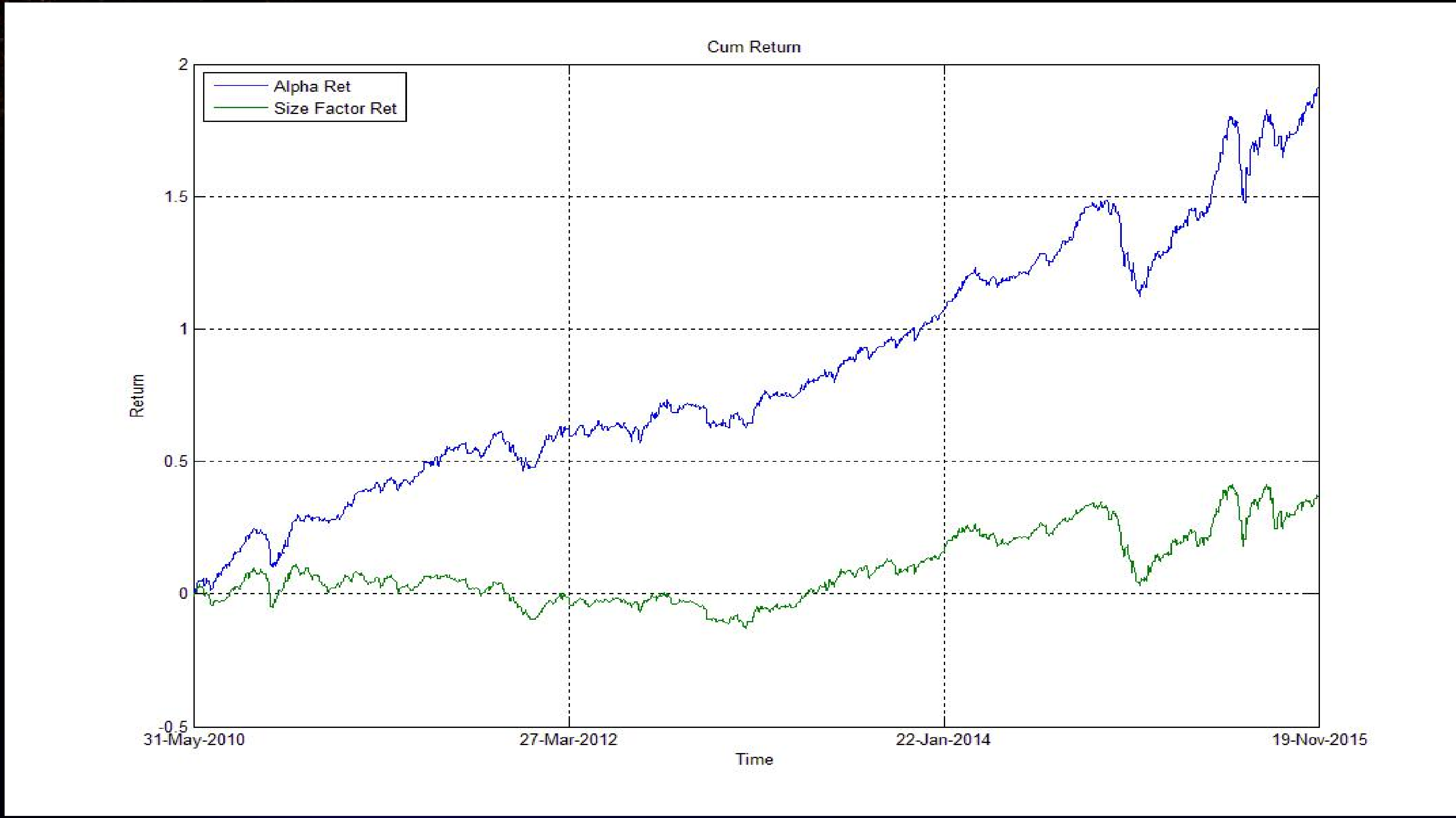

Rongzhi - fixed income index's return and volatility
6 month moving average of rate of return
6 month moving average of volatility of return
Rongzhi - fixed income index


Rongzhi- CTA indexs' Sharpe ratio
Rongzhi- CTA index
Sharpe ratio


Rongzhi - fixed income index's Sharpe ratio
Rongzhi - fixed income index
Sharpe ratio

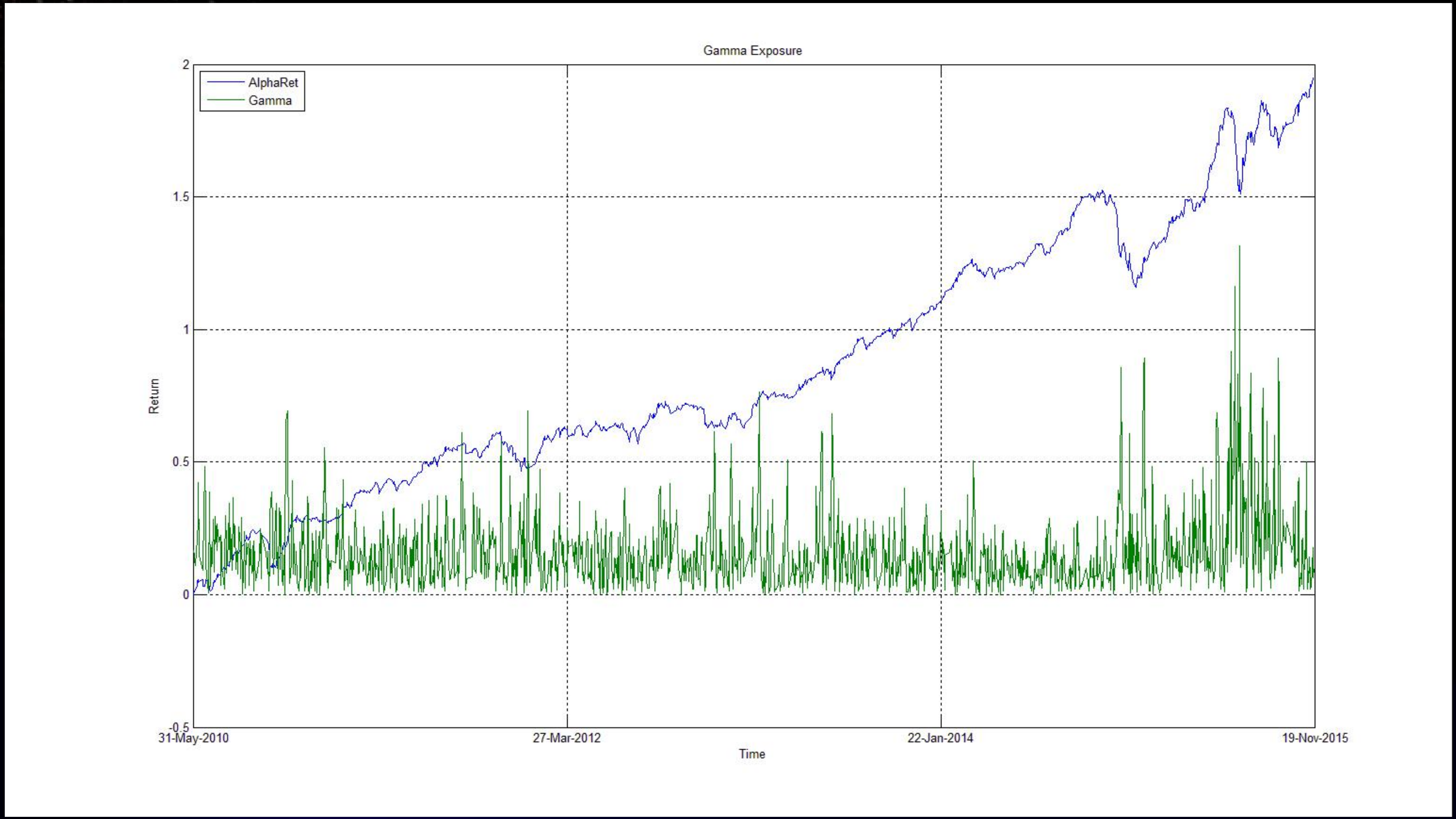
收益归因分析(线性)

- 超额收益有市值因子暴露



市场的情绪（非线性）

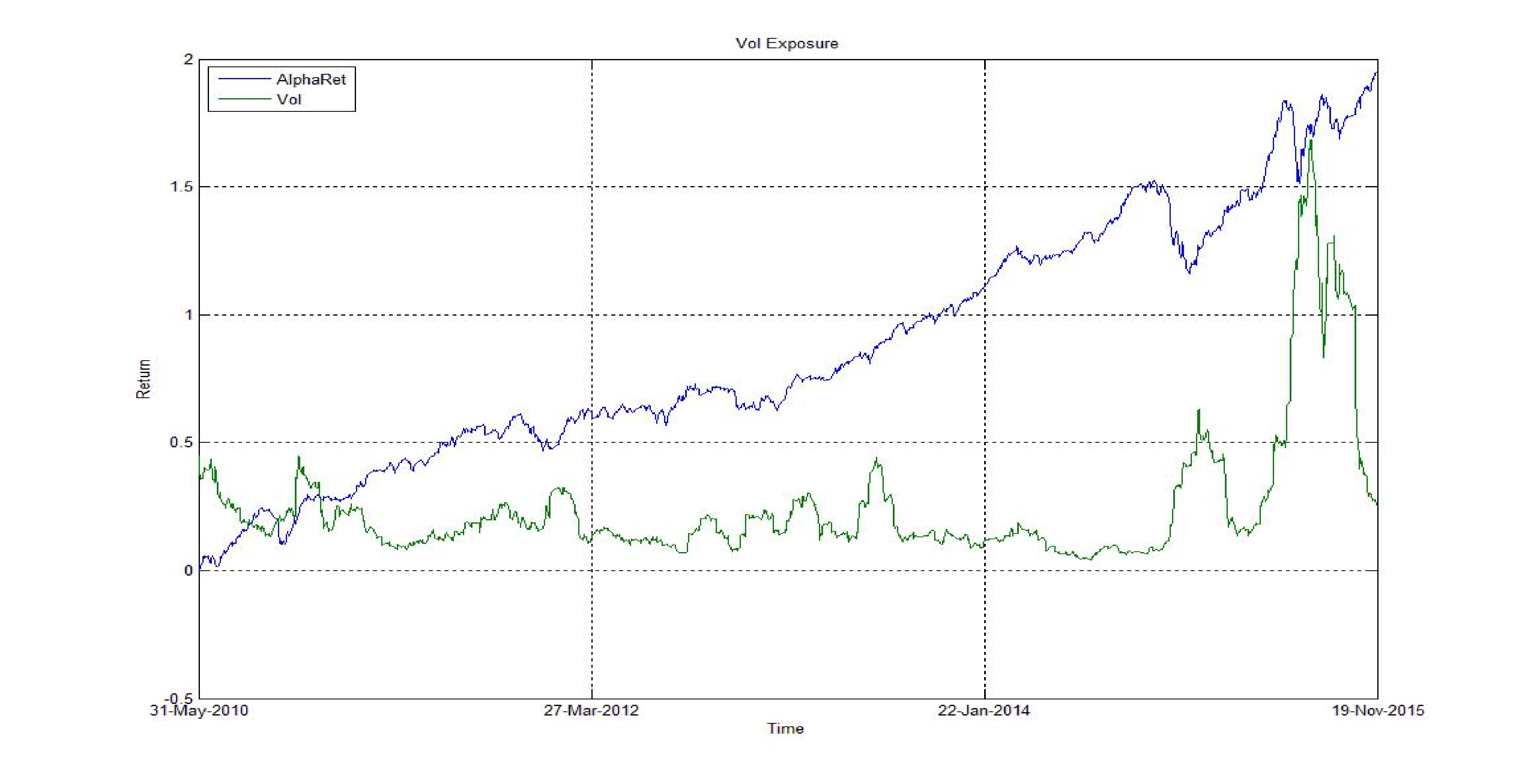
- 超额收益是空gamma的绝对值



清華大學 深圳国际研究生院
Tsinghua Shenzhen International Graduate School

波动性的预测（非线性）

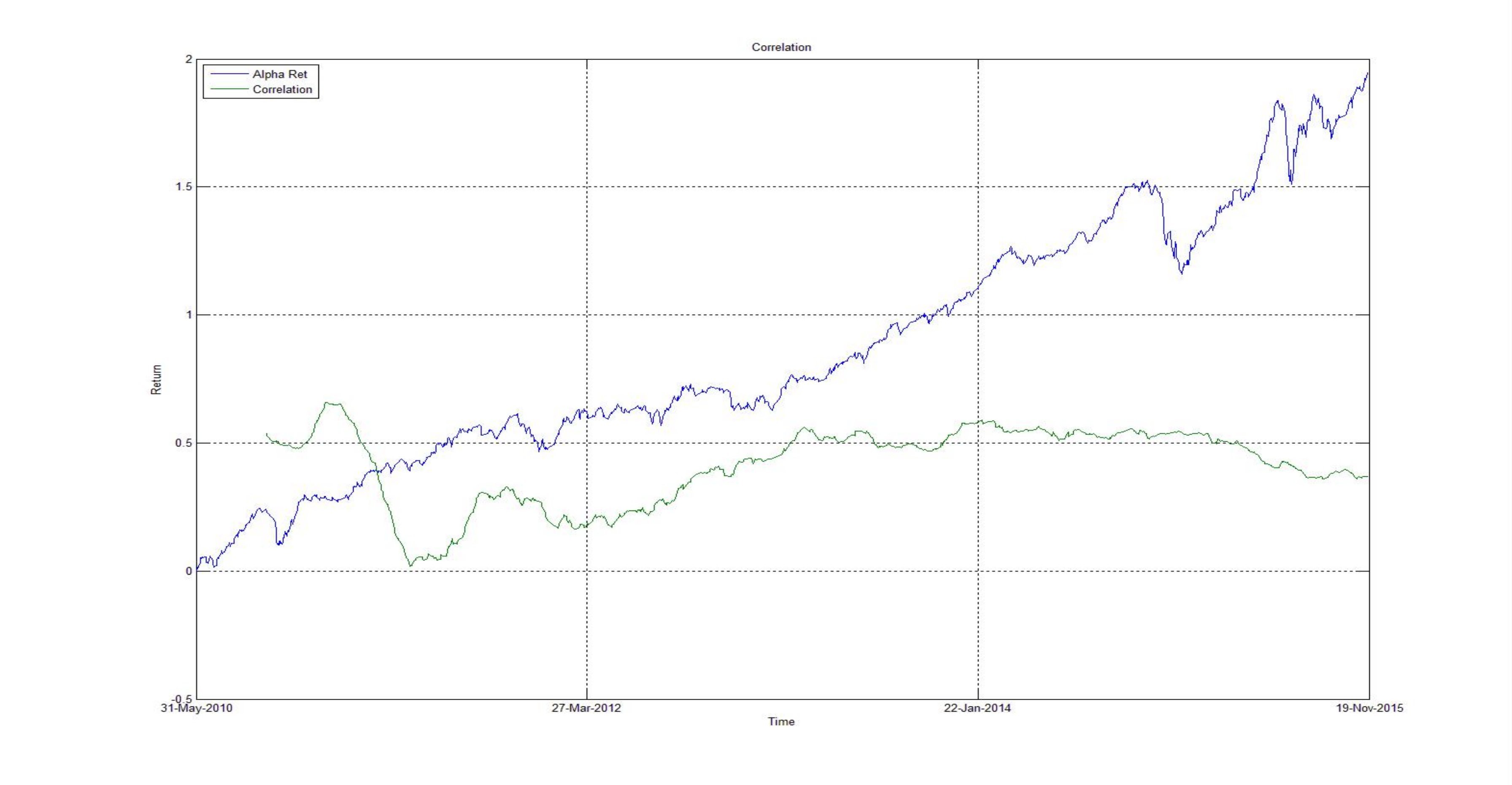
- 套利收益是空波动性



清華大學 深圳国际研究生院
Tsinghua Shenzhen International Graduate School

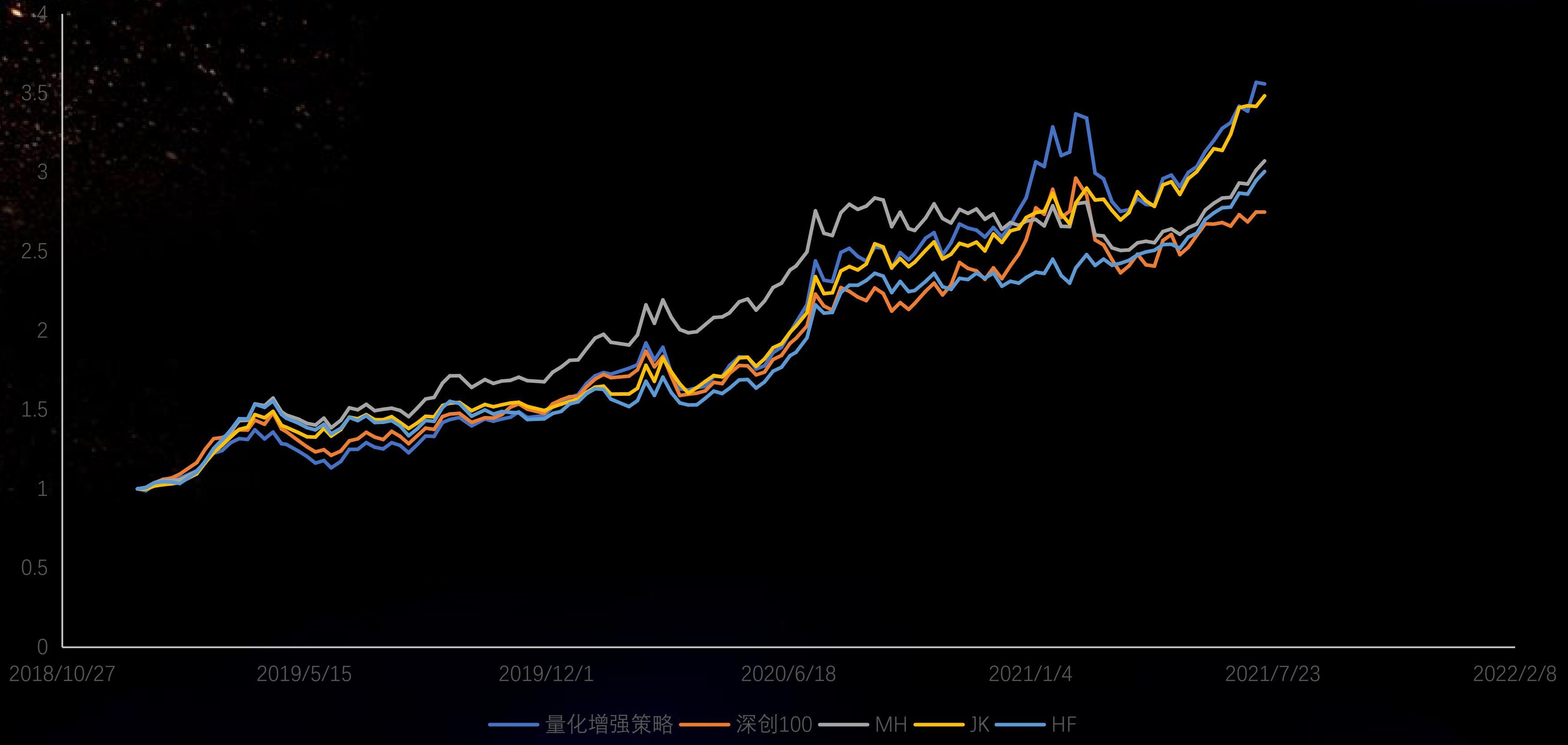
市场相关性的预测（非线性）

- 持续的高相关性会导致超额收益回撤



清華大學深圳国际研究生院
Tsinghua Shenzhen International Graduate School

头部量化私募业绩比较



	量化增强策略	深创100	HF	JK	MH
总收益率	255.77%	174.81%	200.45%	248.25%	207.13%
波动率	23.38%	23.55%	22.90%	23.80%	23.42%

感谢聆听

清華大學 深圳国际研究生院
Tsinghua Shenzhen International Graduate School