Quantifying Trading Behavior in Financial Markets Using Google Trends
使用 Google 趋势量化金融市场中的交易行为

汇报人：严寒、肖宇婷、向紫芊
日　期：2022.9.16

文献基本信息

- **发表期刊**：Nature----Scientific reports（2013）
- **IF**：4.996（Q2）

Tobias Preis

Professor of Behavioural Science at the University of Warwick & Fellow at the Turing Institute
Verified email at tobiaspreis.de

Computational Social Science　Data Science　Machine Learning　Artificial Intelligence　Forecasting

Helen Susannah Moat

Professor of Behavioural Science, University of Warwick; Fellow, The Alan Turing Institute
Verified email at wbs.ac.uk

computational social science　data science　online data　human behaviour　forecasting

H. Eugene Stanley

Professor of Physics, Boston University
Verified email at bu.edu

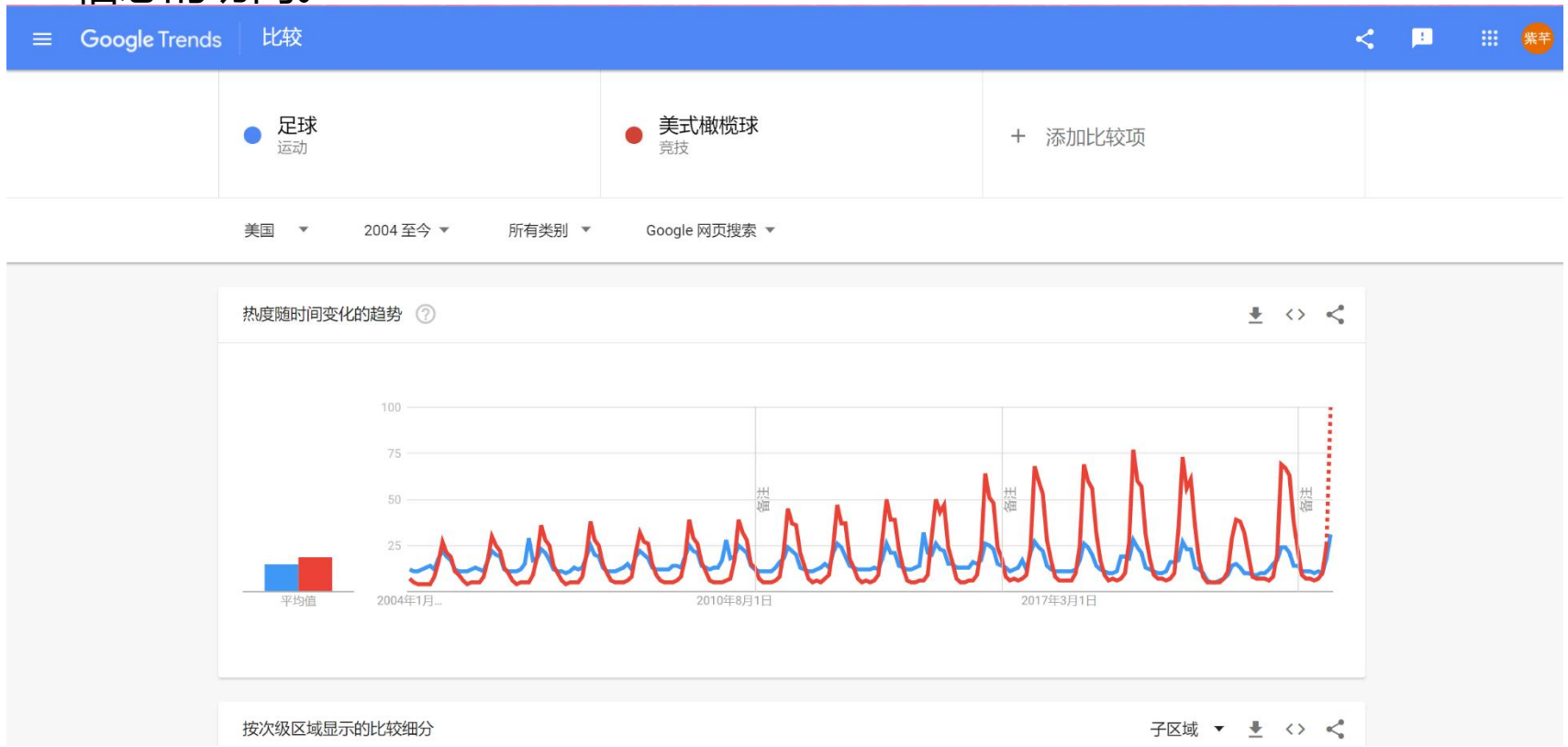
Water　Networks　Econophysics　Biophysics　Complex Systems

引言

引言--研究背景

- **金融行业**：金融行业需要收集信息来做出决策，群体日常活动产生的大数据为量化交易提供了新的信息来源。

- **谷歌趋势（google trends）**：谷歌其提供对不同搜索词的查询量以及这些量如何随时间变化的汇总信息的访问。



研究目的：
量化 Google Trends中**金融术语搜索量的变化与股票价格的关系**，提供对股票交易决策之前的**信息收集过程**的新见解，并以此构建交易策略来进行实现盈利。

引言—相关研究

- 来自**特定国家/地区的搜索结果的点击次数**与**该国家/地区的投资额**相关。(Mondria, J., Wu, T. & Zhang, Y, 2010)

- Google Trends数据的分析表明，选定**搜索词的查询量**变化反映了当前**流感病例数量** (Ginsberg, J.et al, 2009）和**股票市场交易量**的变化 (Choi, H.& Varian, H, 2012)。

- 来自 **Google Trends的数据**可以与各种经济指标的当前值相关联，包括**汽车销售、失业申请、旅游目的地规划和消费者信心。**(Choi, H.& Varian, H, 2012)

- **股票市场交易量**和**搜索量**之间的联系使用 Yahoo! 进行了复现。(Bordino,1.et al, 2012)

- 来**自人均 GDP 较高国家的互联网用户**比过去几年更有可能**搜索有关未来年份的信息。**(Preis, T., Moat, H. $.,2012)

引言—研究假设

- 2004年至2011年期间，Google Trends某些**金融术语的搜索查询量**可以用于**构建交易策略，这些数据不仅反映了股票市场的当前状态，而且还能够预测某些未来趋势。**

- 在**担忧期**（金融市场的显著下跌之前），投资者可能会**寻找更多关于市场的信息**，具体表现在某些**金融相关的术语的搜索量增加**，然后再决定买入或卖出，反之同理。具体表现如下：

  - 股票价格**上涨**之前某些金融相关术语的搜索量**下降。**
  - 股票价格**下跌**之前某些金融相关术语的搜索量**增加。**

- 在 2004 年到 2011 年期间，谷歌趋势中某些**金融术语搜索量的变化**可以**预测更有利的交易策略 。**

研究方法

研究方法—Google Trends strategy（谷歌趋势投资策略）

- 为了研究谷歌趋势数据捕获的**信息收集行为**的变化是否与 2004 年至 2011 年期间**股价的后期变化**有关，作者设计了以下投资策略（**Google Trends strategy**）来模拟投资行为，分别为：

  - 信息收集行为**增加**时，$\Delta n(t-1,\Delta t)>0$, 在第 t 周的第一个交易日以收盘价 p(t) 卖出 DJIA，并在下周第一个交易日以价格p(t+1）买入。（**即空头，从市场价格下跌中获利，之前的假设：搜索量增加时，意味着经济形势不好，交易者会在做出决策前在网上搜寻相关信息**）。

  - 信息收集行为**降低**时，$\Delta n(t-1,\Delta t)<0$,在第 t 周的第一个交易日以收盘价 p(t) 买入 DJIA，并在下周第一个交易日以 p(t + 1) 的价格卖出 。（**即多头，从市场价格上涨获利，预期股价会上涨，经济形势较好**）

研究方法—Google Trends strategy（谷歌趋势投资策略）

- 为了研究谷歌趋势数据捕获的**信息收集行为**的变化是否与 2004 年至 2011 年期间**股价的后期变化**有关，作者设计了以下投资策略（**Google Trends strategy**）来模拟投资行为，分别为：

  - 信息收集行为**增加**时，$\Delta n(t-1,\Delta t) > 0$,在第 t 周的第一个交易日以收盘价 p(t) 卖出 DJIA，并在下周第一个交易日以价格p(t+1）买入。（**即空头，从市场价格下跌中获利，之前的假设：搜索量增加时，意味着经济形势不好，交易者会在做出决策前在网上搜寻相关信息**）。

  - 信息收集行为**降低**时，$\Delta n(t-1,\Delta t) < 0$, 在第 t 周的第一个交易日以收盘价 p(t) 买入 DJIA，并在下周第一个交易日以 p(t + 1) 的价格卖出 。（**即多头，从市场价格上涨获利，预期股价会上涨，经济形势较好**）

- 使用**搜索量的相对变化**来量化（t 以周为单位测量）**信息收集行为**：

$$\Delta n(\boldsymbol{t},\Delta \boldsymbol{t}) = \boldsymbol{n(t)} - \boldsymbol{N(t-1,\Delta t)}$$

第t周的搜索量与前Δt周总搜索量平均数的差

其中$\boldsymbol{n(t)}$：在第 $\boldsymbol{t}$ 周内针对特定搜索词（例如：debt债务）进行了多少次搜索；

$$\boldsymbol{N(t-1,\Delta t)} = \frac{\boldsymbol{n(t-1)+n(t-2)+\ldots+(t-\Delta t)}}{\boldsymbol{\Delta t}}$$

研究方法—定义累积回报R

- 如果持有“空头”，**以收盘价 p(t) 卖出并以 p(t + 1) 价格回购**，那么累积回报R会改变:

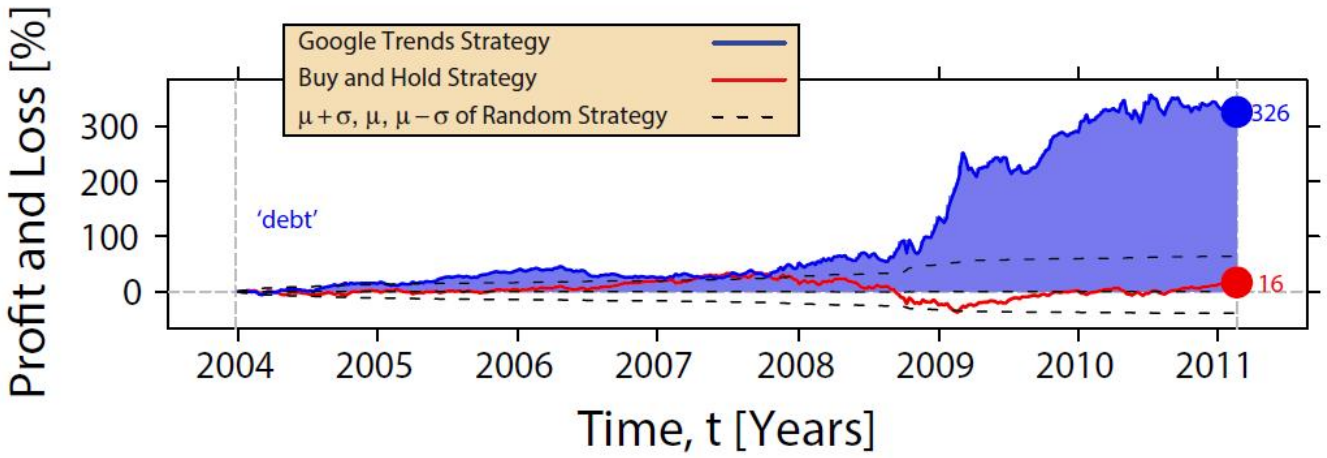
$$\Delta_R = log\big(p(t)\big) - \log(p(t+1))$$

- 如果持有“多头”，**以收盘价 p(t) 买入并以 p(t + 1) 价格卖出**，那么累积回报R会改变:

$$\Delta_R = log\big(p(t+1)\big) - \log(p(t))$$

备注：对数函数在其定义域内是单调增函数，取对数后不会改变数据的相对关系，缩小数据的绝对数值，方便计算。

研究方法—对照组



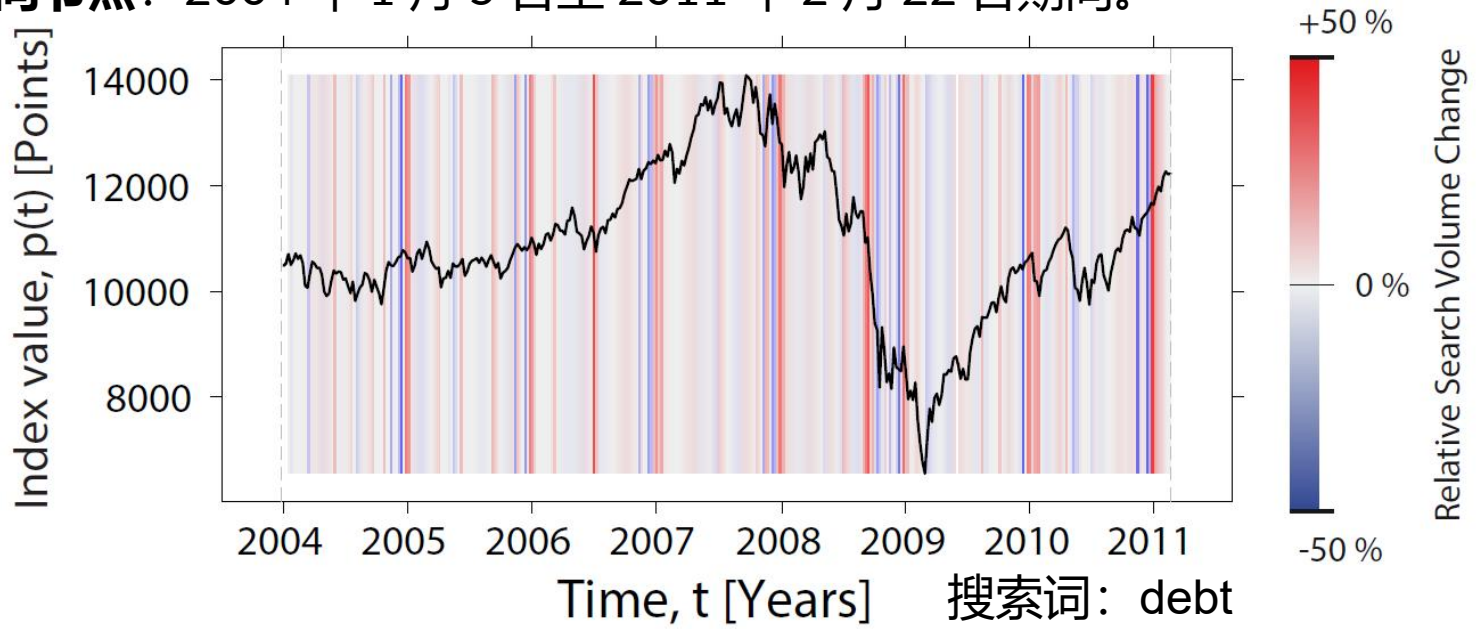
- **随机投资策略**：表示以不相关的随机方式买卖市场指数的策略，标准差来自对随机投资策略的 10,000 次独立实现的模拟。

- **买入持有策略（BUY AND HOLD）：开始时买入指数，然后在持有期结束时卖出**，是一种简单的长期投资策略，**忽略短期金融市场的波动及衰退，专注于长期成长与收益率。**

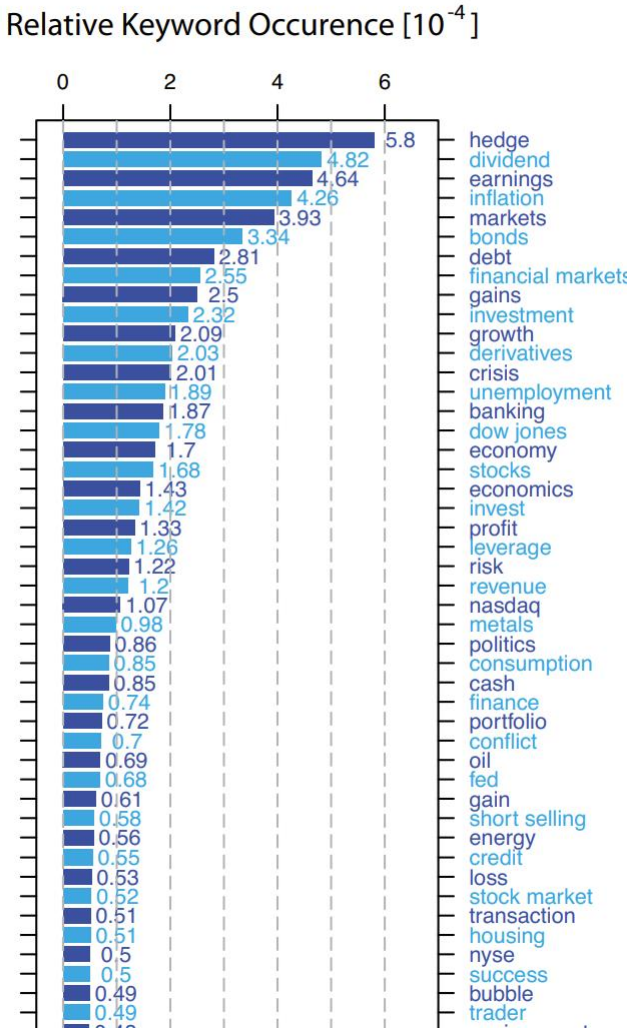
- **道琼斯策略（DOW JONES）：**使用道琼斯工业平均指数 (DJIA) 的每周收盘价$p(t)$预测走势。

研究方法—数据集

- **金融相关术语的确定**
  - **Google Set**：从一组事物集合中输入一些词组， Google Sets 将尝试预测集合中的其他词。
  - 最后选取了**98**个金融相关术语。
- **金融相关术语搜索量的收集（Google Trends)**
  - **n(t – 1)**：在第 t – 1 周内针对**特定搜索词**（例如：debt债务）进行了多少次**搜索** 。
- **道琼斯工业平均指数 (DJIA)**
  - **定义**：美国证券交易所上市的30家著名公司价格加权平均值， p(t)表示在第 t 周的第一个交易日的收盘价。
- **数据收集时间节点**：2004 年 1 月 5 日至 2011 年 2 月 22 日期间。



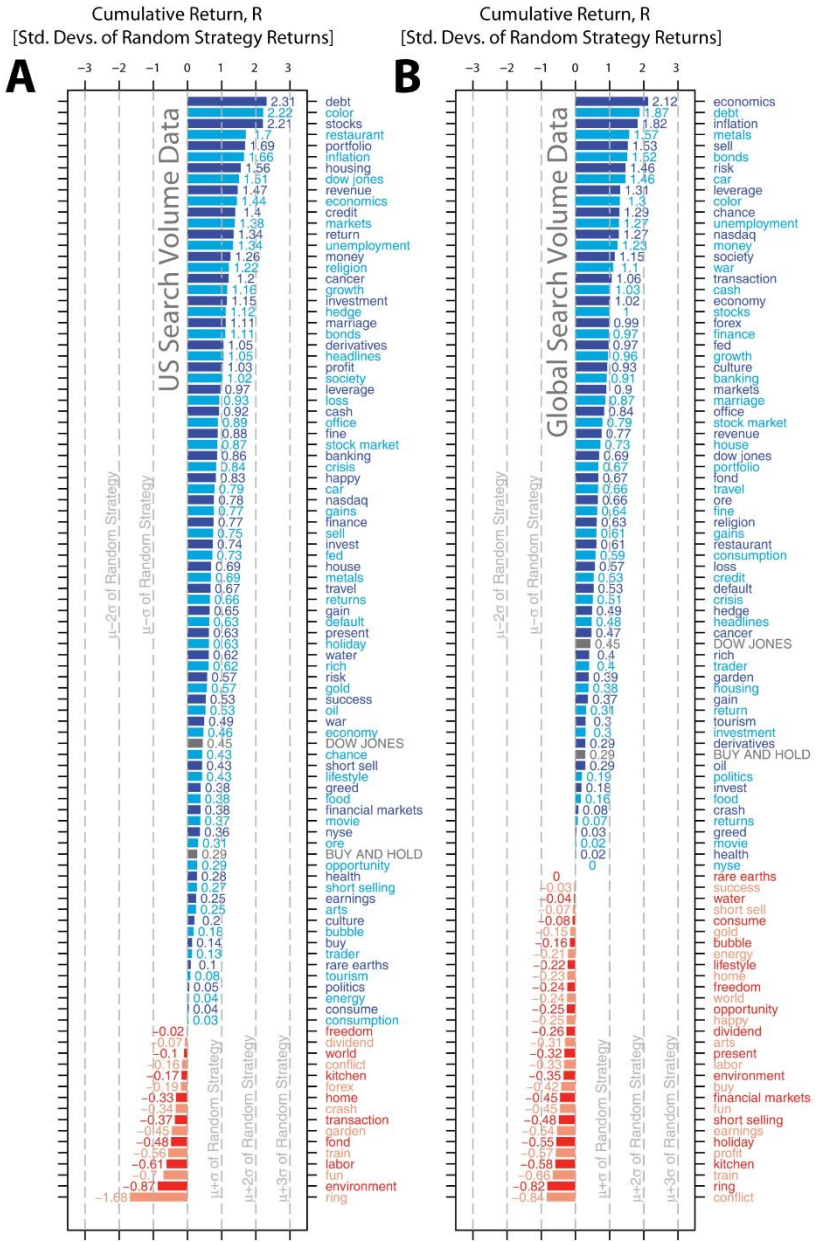
研究方法—术语金融相关性的计算



- **术语金融相关性的计算**
  - 通过计算 2004 年 8 月至 2011 年 6 月**金融时报**中每个搜索词的频率来量化金融相关性
  - 通过每个搜索词的**谷歌点击次数**对术语金融相关性进行**标准化**。

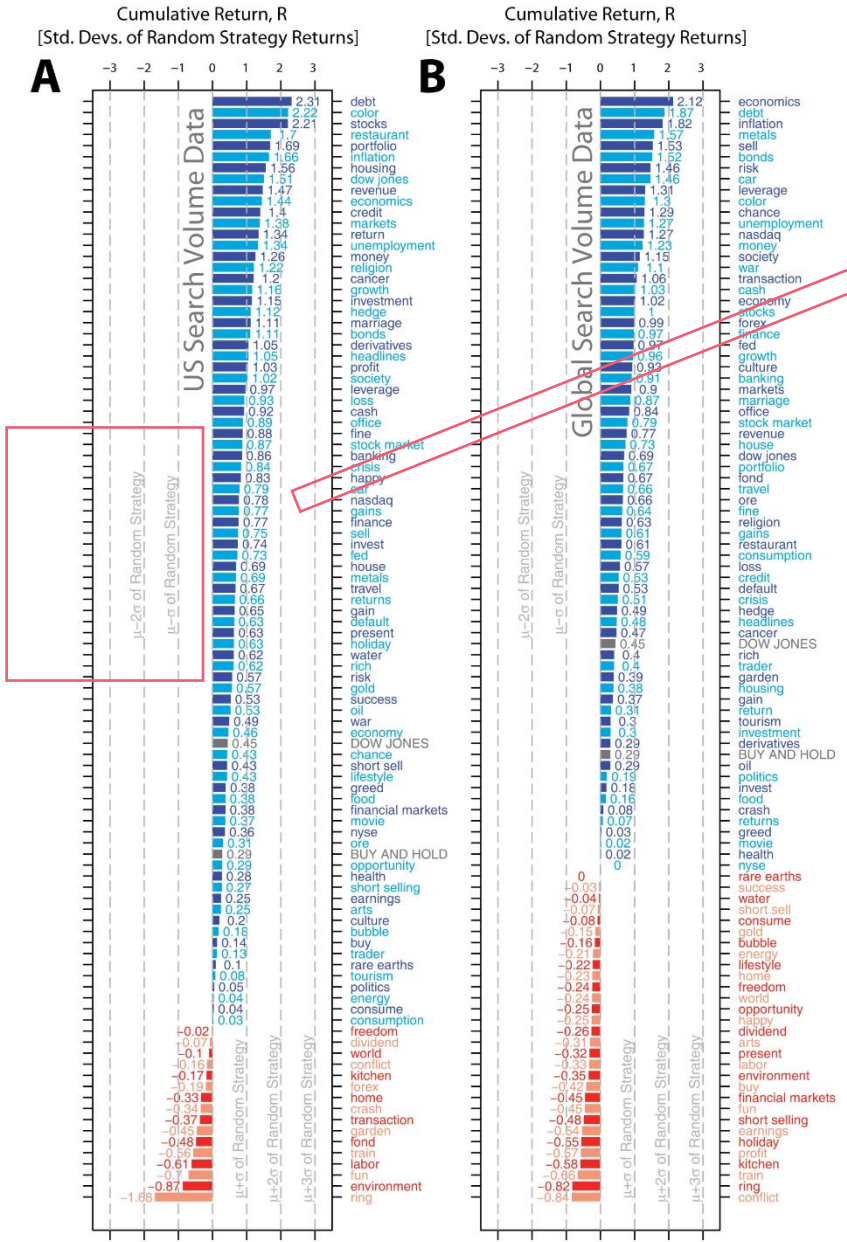
数据分析

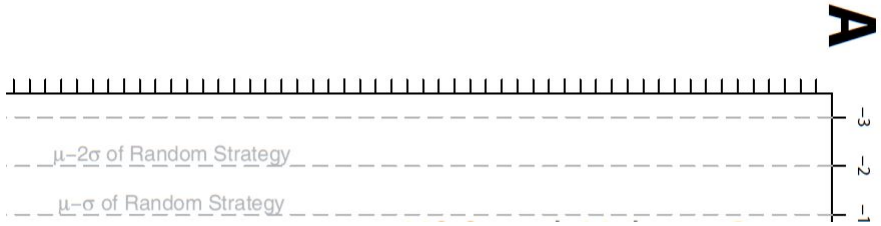
数据分析



- 在交易开始时，将所有投资组合的价值设置为**1**。
- 为了确保结果的稳健性，基于给定搜索词的策略的整体交易表现被确定为$\Delta t = 1...6$ **周获得的六次回报的平均值。**
- 根据交易表现（即累计回报R）对98个搜索词进行排名，其中图A：仅使用美国用户的搜索数据；图B：全球生成的搜索量。
- **蓝色**表示**正回报**，使用两种**红色**表示**负回报**
- **虚线**对应随机策略的 -3、-2、-1、0、-1、-2、-3 个标准差。

数据分析—随机投资策略

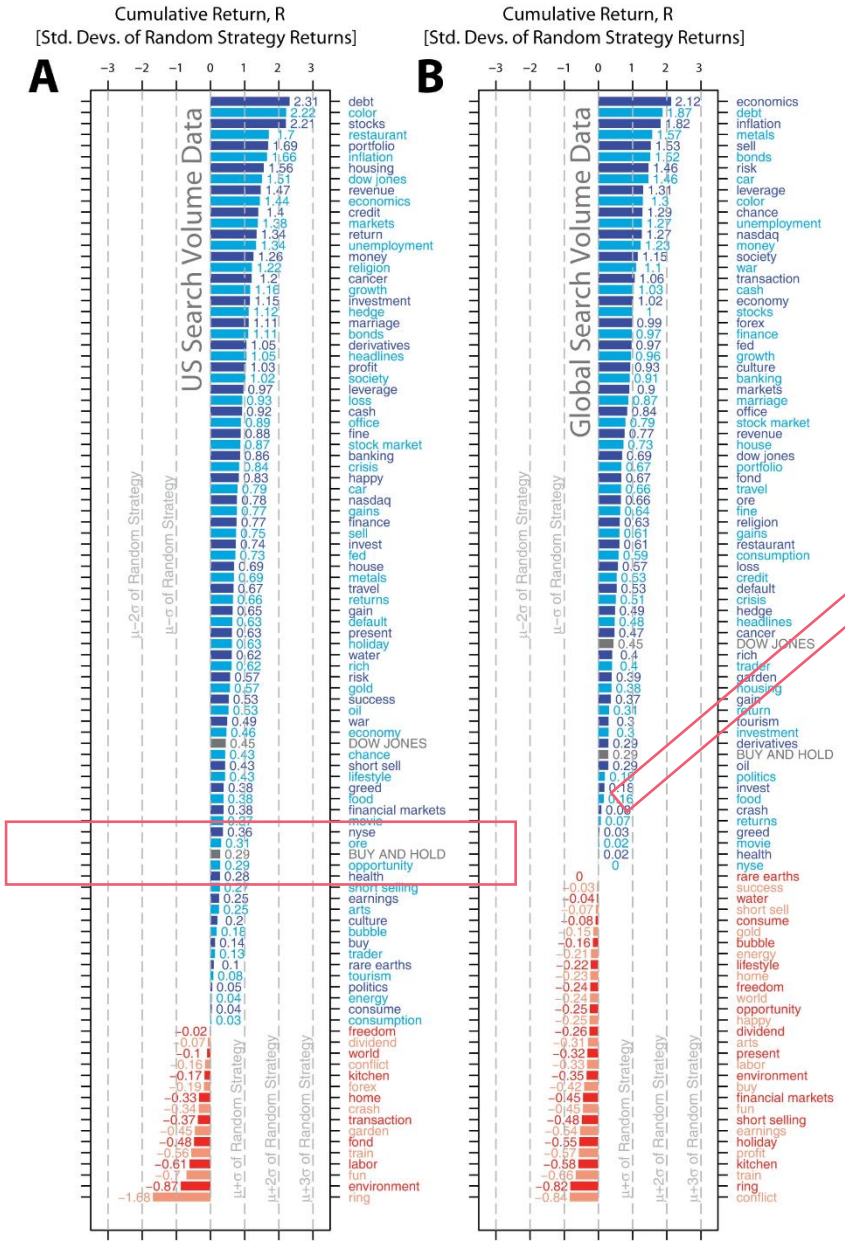


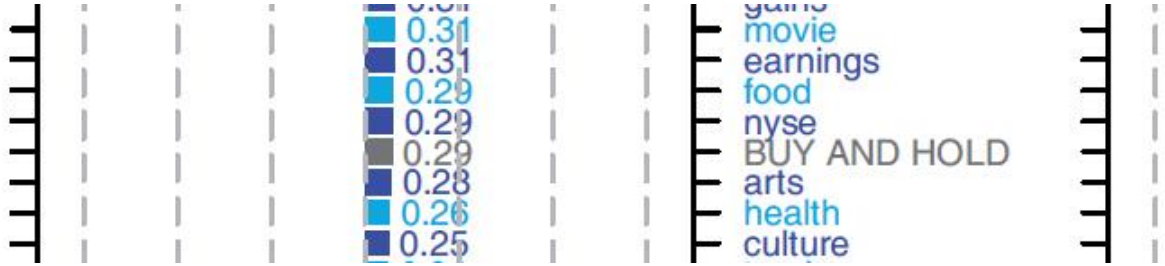


随机投资策略

- **随机投资策略**产生的最终投资组合价值分布接近**对数正态分布。**
- 从这些投资组合值的对数得出的**随机投资策略的累积回报也服从正态分布**，平均回报率R=0 。

数据分析—BUY AND HOLD(买入并持有策略)

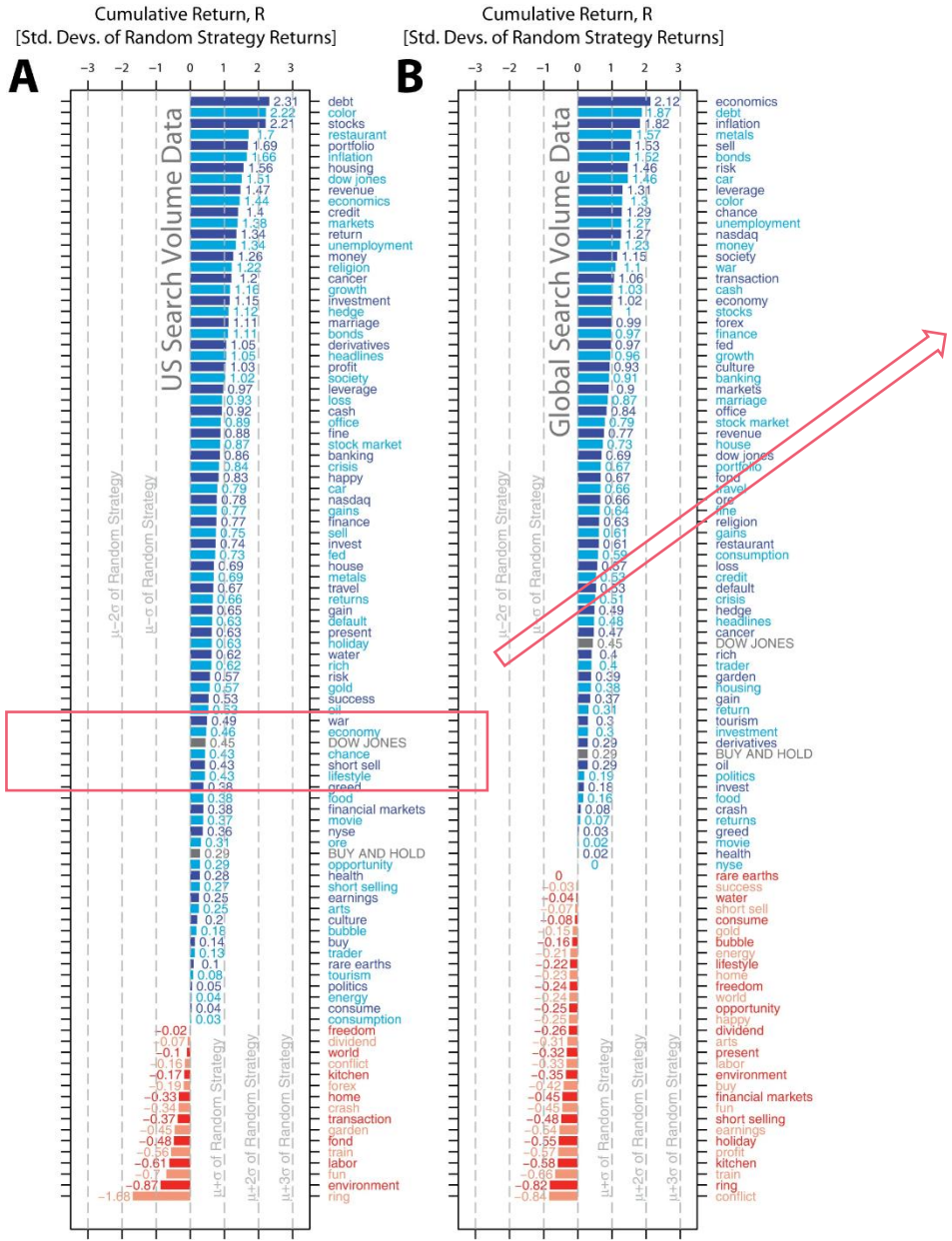




BUY AND HOLD(买入并持有策略)

- 在开始时买入并在持有期结束时卖出，该策略产生 **16%** 的利润，相当于 DJIA 在 2004 年 1 月至 2011 年 2 月期间的整体价值增长。

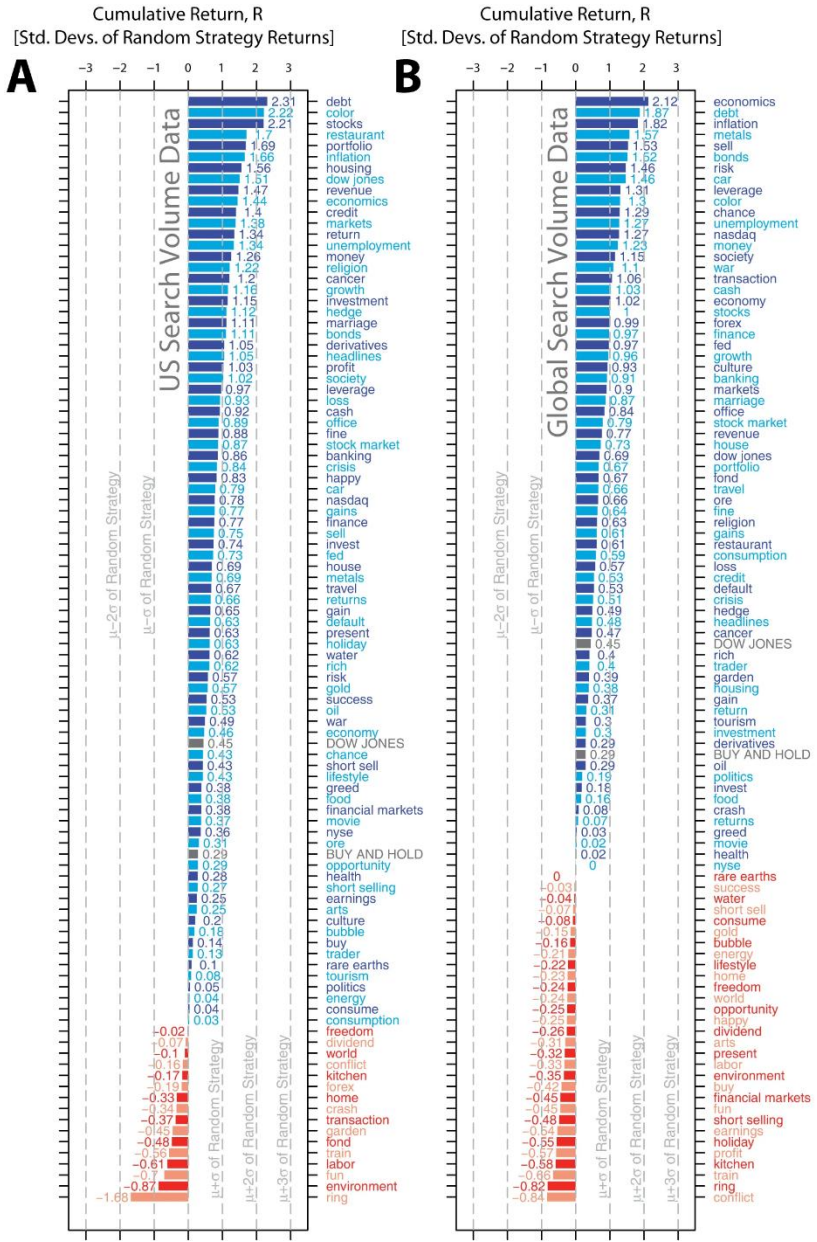
数据分析—DOW JONES(道琼斯策略)



DOW JONES(道琼斯策略)

- 使用道琼斯工业平均指数 (DJIA) 的每周收盘价预测走势。
- 该策略在 Δt = 3 周时产生 33% 的利润。
- 计算Δt = 1...6 周获得的六个回报的平均值时，**R 为不相关随机投资策略的累积回报的 0.45 个标准差。**

数据分析—不同策略比较



- **不论是美国还是全球，Google Trends策略的总体回报率明显高于随机策略的回报率。**（<R>US= 0.60;t = 8.65, df = 97,p < 0.001,one sample t-test；<R>Global= 0.43;t = 6.40,df = 97,p < 0.001,one sample t-test)

- 搜索量分析的范围扩展到全球用户会降低谷歌趋势交易策略在美国市场获得的总体回报，**基于全球搜索量数据的策略不如基于美国搜索量数据的策略。**(<R>US= 0.60,<R>Global= 0.43;t = 2.69, df = 97, p < 0.01,two-sided paired t-test)

- Google Trends策略的效果因所选的搜索字词而异，**搜索词相关的回报与其金融相关性相关。**(Kendall’s tau = 0.275,z = 4.01，N = 98,p < 0.001)

数据分析—结论

- **实证结果与两部分假设一致：**
  - 道琼斯工业平均指数价格的关键**上涨**之前，某些与金融相关的术语的搜索量**减少。**
  - 道琼斯工业平均指数价格的关键**下降**之前，某些与金融相关的术语的搜索量**增加。**

- **交易策略可以分解为两个策略组成部分：**
  - 搜索量的**减少**促使**买入**
  - 搜索量的**增加**促使**卖出**

数据分析—验证结论

- **策略一：搜索量减少时只买入**
  - 回报**显著高于**随机投资策略。(<R>USLong = 0.41; t = 11.42, df = 97, **p < 0.001**, one sample t-test)
  - 搜索词的金融相关性与回报之间存在**正相关**关系。(Kendall's tau = 0.242, z = 3.53, N = 98, **p < 0.001**)
- **策略二：搜索量增加时只卖出**
  - 回报**显著高于**随机投资策略的回报。(<R>USShort = 0.19; t = 5.28, df = 97, **p < 0.001**, one sample t-test)
  - 搜索词的金融相关性与回报之间存在**正相关**关系。(Kendall's tau = 0.275, z = 4.01, N = 98, **p < 0.001**)

讨 论

讨论—研究结论

- 结果与假设一致，**谷歌趋势数据不仅反映了当前经济状况的各个方面，而且可能还提供了对经济行为者行为未来趋势的一些洞察。**
- 使用 2004 年 1 月至 2011 年 2 月期间的历史数据，发现在股市**下跌之前**，谷歌搜索与金融市场相关的关键字的**搜索量有所增加。** **表明，搜索量数据中的这些警告信号可以被用于构建有利的交易策略。**
- 基于美国用户搜索量数据的策略在美国市场上比使用全球搜索量数据的策略更成功。
- 将金融交易数据等大型行为数据集与搜索查询量数据相结合，**可能会为大规模集体决策的不同阶段提供新的见解**，进一步说明了**新的大数据集提供的令人兴奋的可能性，以促进对社会中复杂集体行为的理解。**

讨论

- 可能原因
  - **Herbert Simon** 的决策模型（**有限理性模型**） 的背景下对结果提供了一种可能的解释。 谷歌趋势数据和股市数据可能反映投资者决策过程中的后续阶段。 在金融市场上**以较低价格出售的趋势可能会出现一段时间的担忧**，在这样的担忧时期，人们可能**倾向于收集更多关于市场状况的信息**，这种行为可能反映在谷歌趋势金融相关性术语**搜索量的增加**上。
- 未来展望
  - 在这项工作中，作者**量化了搜索量变化与股票市场价格变化之间的关系。** 未来的工作将需要对**导致人们在以较低价格出售股票之前搜索债务等术语的潜在心理机制进行解释。**

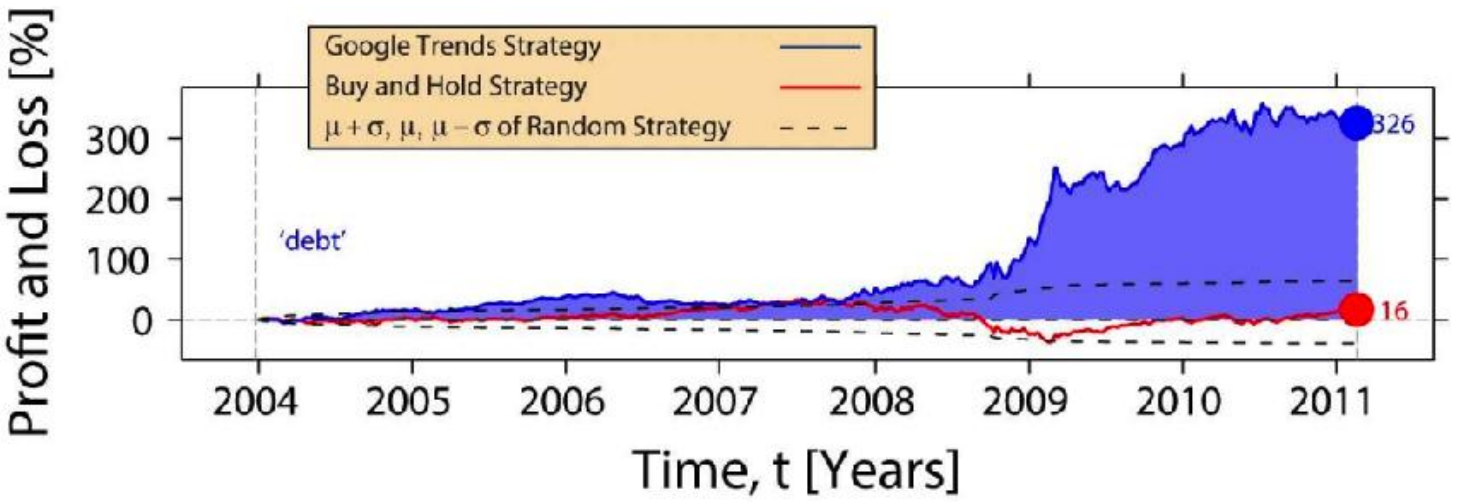
文献评述

文献评述

- 基于与某一领域或群体现象相关的关键词，在某一时间内大量的出现或减少，其潜在的原因可能和群体活动和社会现象相关，通过大数据方法来进行主题分析并结合一些现象，可以预测未来和反映当下，但还是需要**结合理论和实际**，不能只凭数据来剖析人的心理活动。

- 该研究分析的是2004年至2011年期间的数据，2007年至2011年，**美国次贷危机及全球金融危机**这个时代背景下，是否会产生**极端值**而影响了结果？在2004年至2011年期间，随着网络的普及，**网民的数量是呈增长趋势的**，其上网的频率和搜索的频率也可能是逐年增长的，随着时间和科技发展的变化，数据搜索量也在变化。

文献评述



- 自2008年以后，用“debt”一词的搜索量的谷歌趋势策略其利损值显著的增长，在2008年前却没有这种显著变化，且补充资料中的大部分与金融相关的术语都是在2008年以后有显著的利润变化。

- 所以是否是因为**股民上网人数**增加使搜索量增加而导致的利损值显著的增长，还是因为正在经历**全球危机**使“debt”一词搜索量增加导致的预测的利润量增加呢？又或许是两者的相互作用？这一点并没有在文章中体现，其结果可能会存在偏差，可能不具有大的外部效度。

请老师批评指正