Complex model to complex data

Feng Li
<feng.li@cufe.edu.cn>

School of Statistics and Mathematics
Central University of Finance and Economics

http://feng.li/

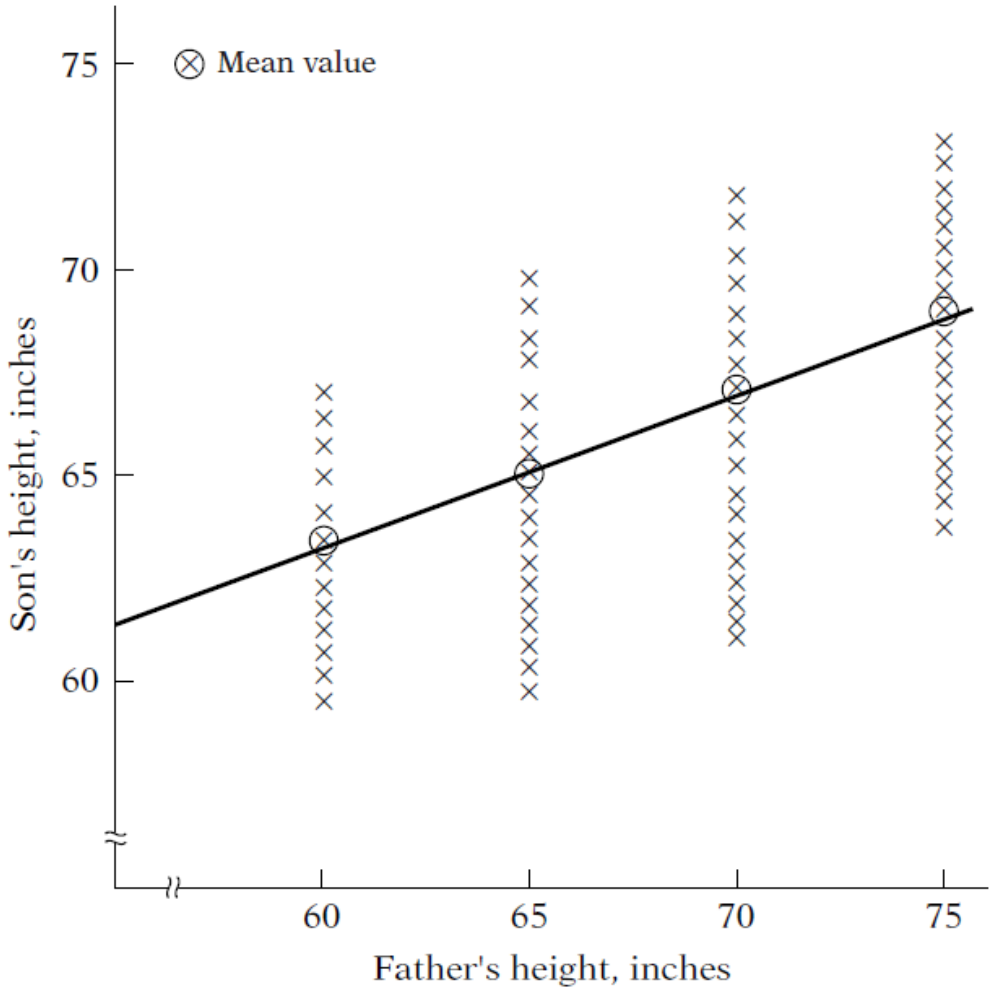
Have you ever thought about...

- Why the weather forecast is not accurate sometimes?
  (rain or not $\Leftrightarrow$ cloudy, humidity, historical data)
- How does the email filter know whether a mail is a spam or not?
  (spam or not $\Leftrightarrow$ sender, keywords)
- Can we predict the next financial crisis?
  (Next crisis time $\Leftrightarrow$ when was last time, stock prices, exchange rate, unemployment rate)
- ...

Statistical modeling is trying to formalize (**model**) the relationships among the variables (**data**).

Toy examples

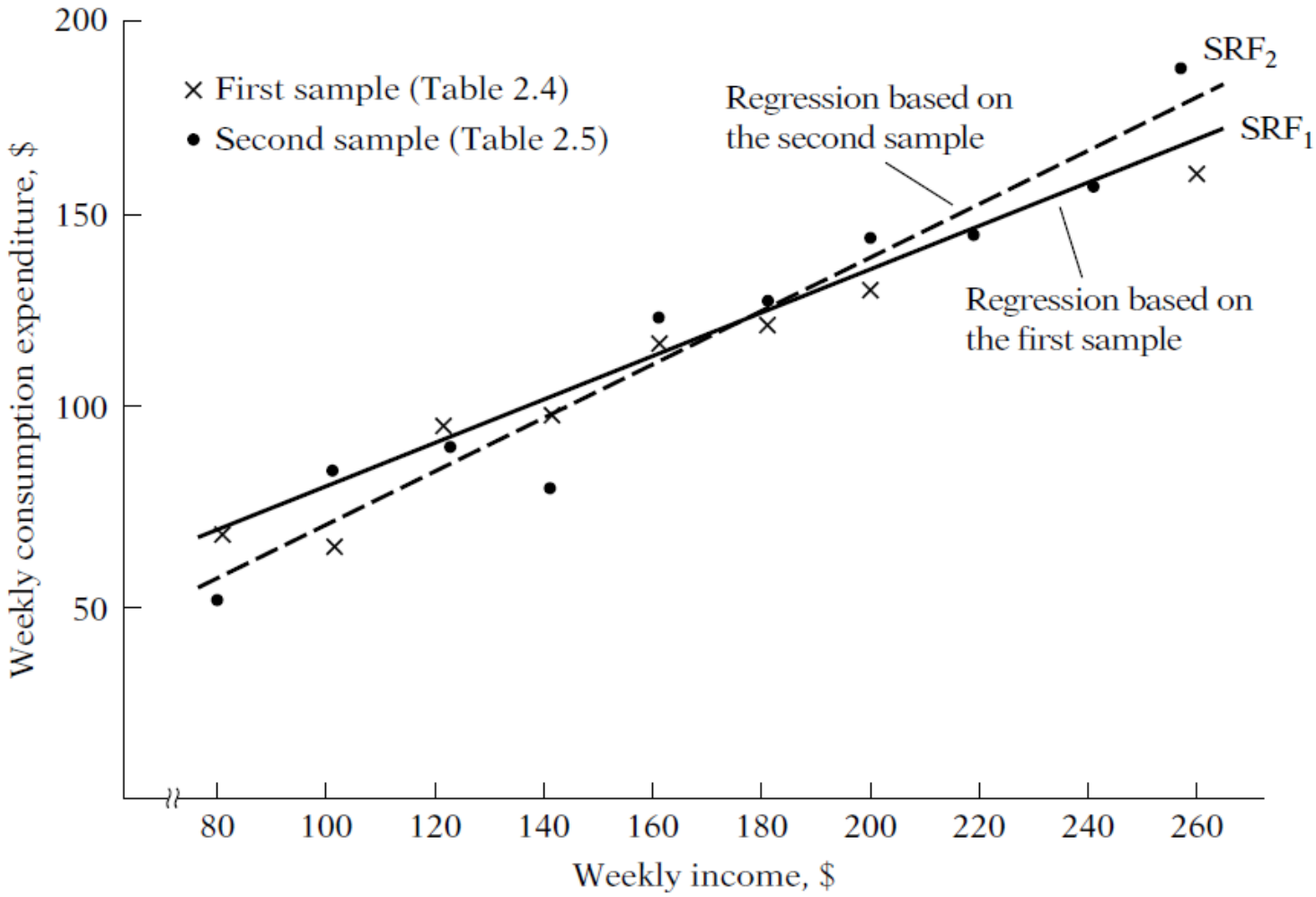
↛ Father's height vs son's height



Gujarati, D. N. (2003). Basic Econometrics. 4th.

Toy examples

↛ Family income and consumption

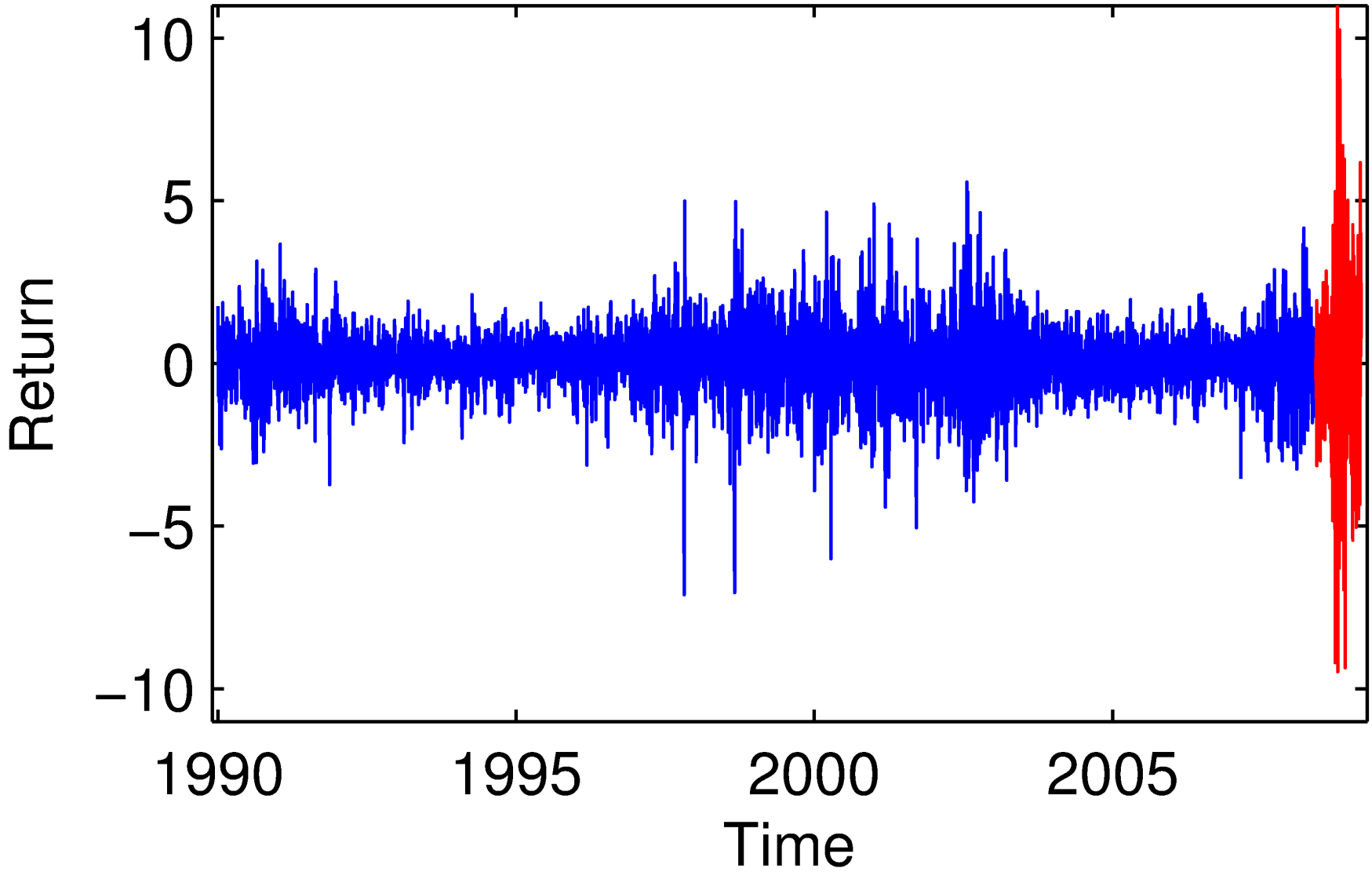


Gujarati, D. N. (2003). Basic Econometrics. 4th.

- Models are simple.
- Works well at most situations.
- Easy to imagine and implement.
- It takes less than 1 second to have the result with a laptop.

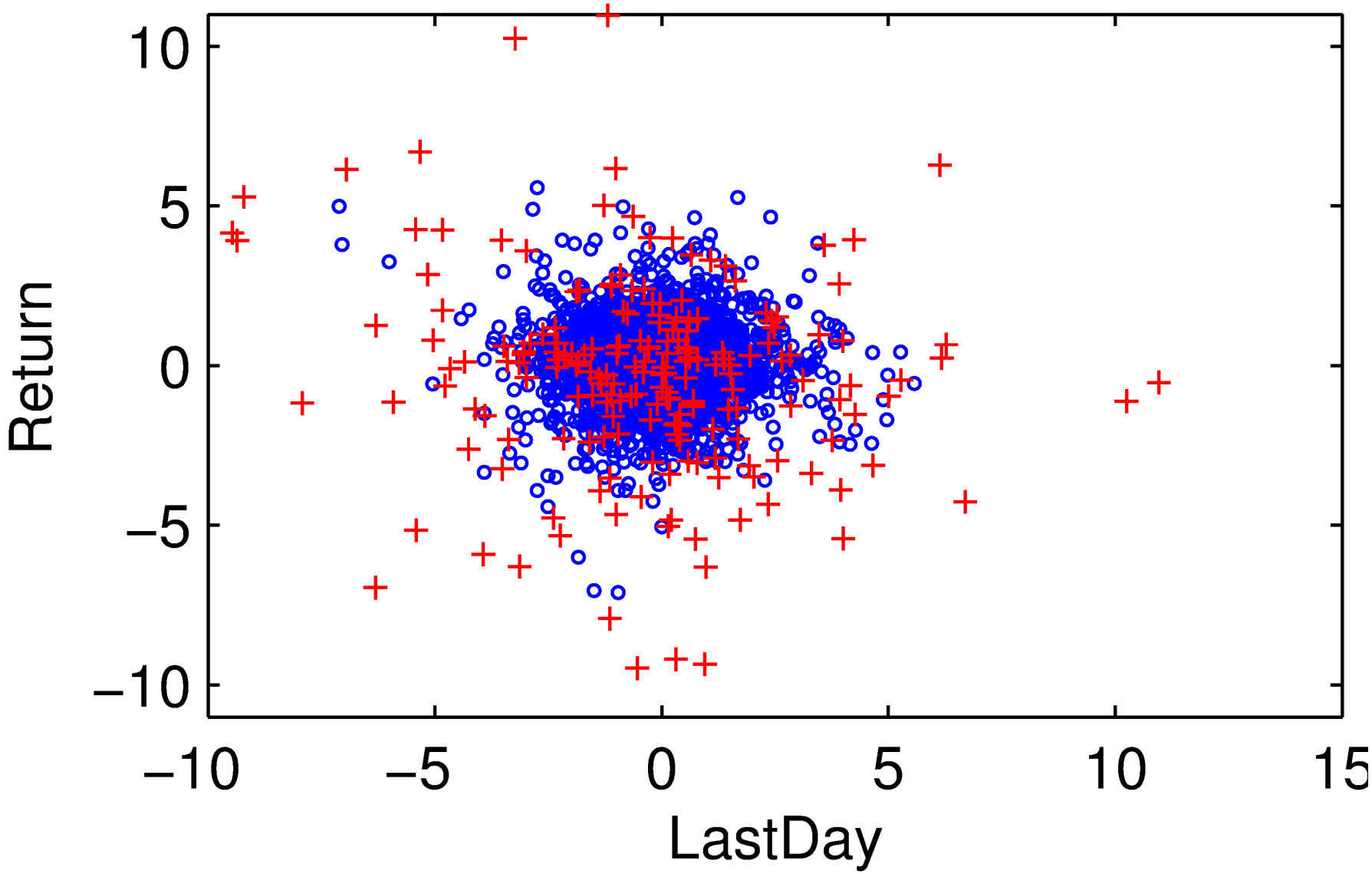
The typical data in finance

↛ Daily stock market returns



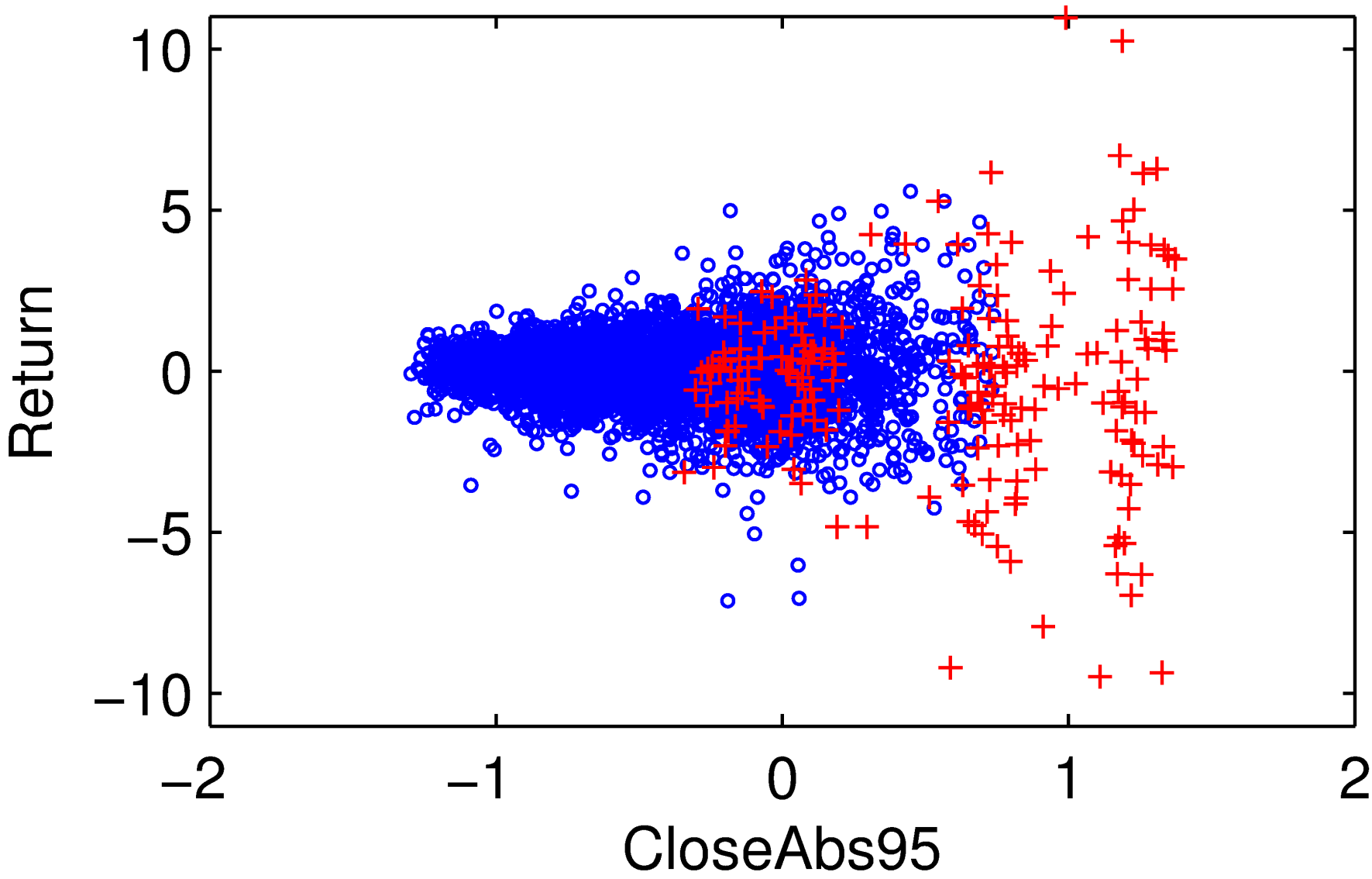
The typical data in finance

↛ Daily stock market returns, a closer look



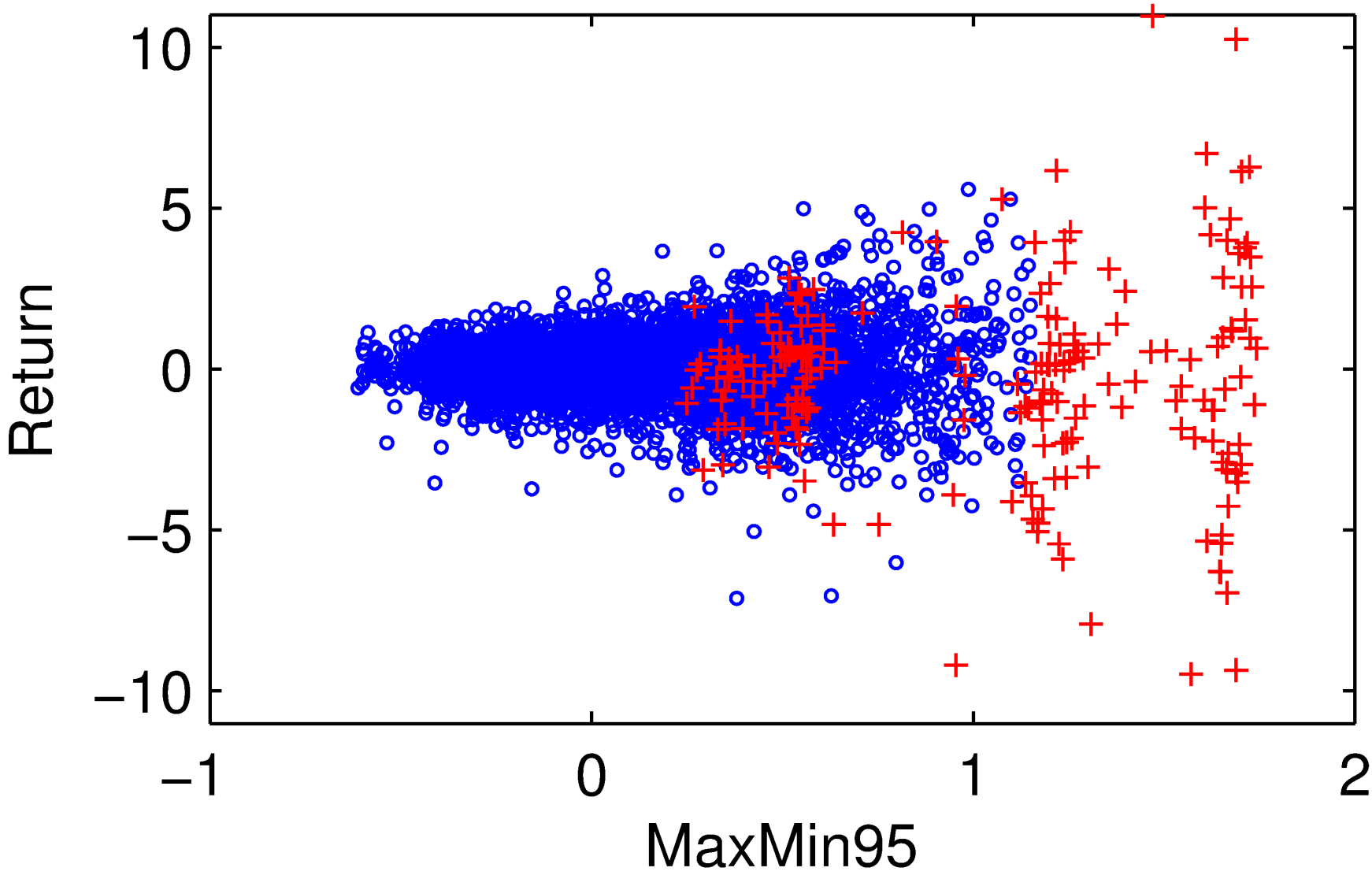
The typical data in finance

↛ Daily stock market returns, a closer look



The typical data in finance

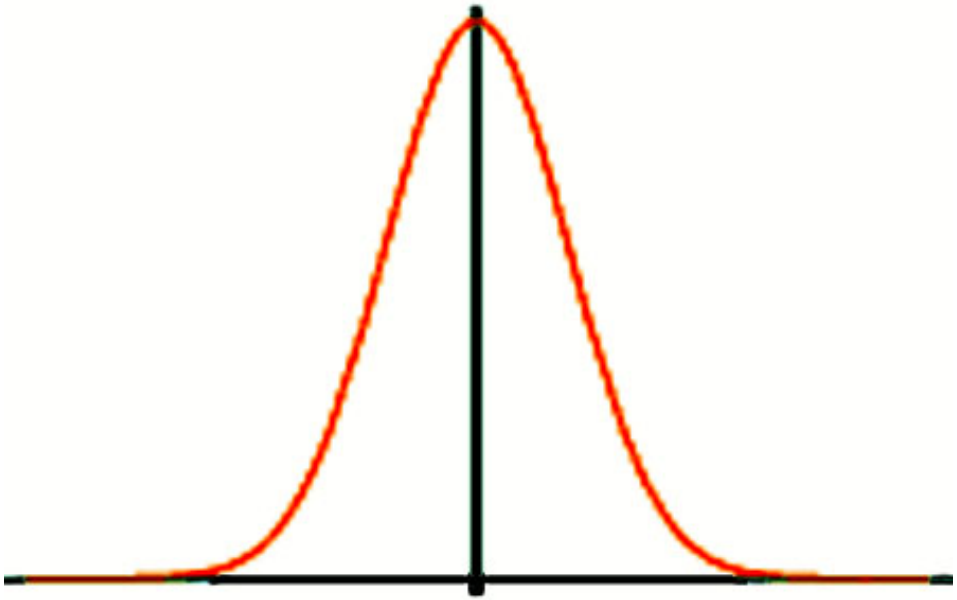
↛ Daily stock market returns, a closer look



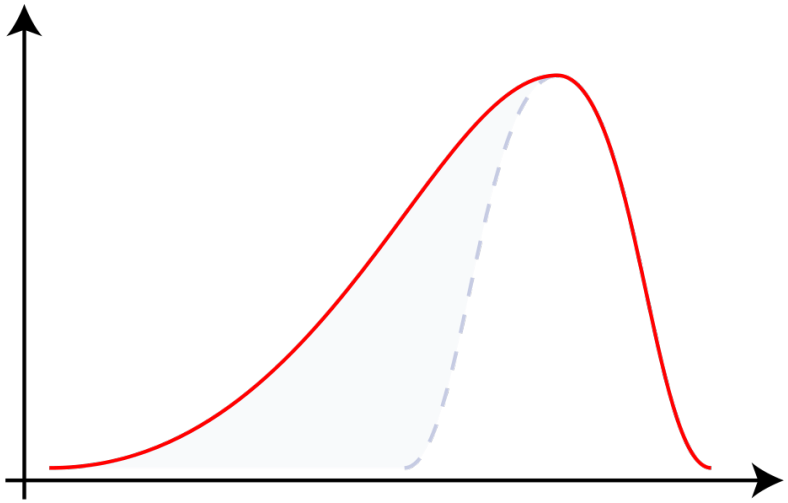
What can we find?

- This does not look like as **normal** (think about the mean and variation)!
- How do we describe it in the language of statistics?
  - We use **mean** and **variance** (*standard deviation*) to describe normality.
  - We use **skewness**, and **kurtosis** (*degrees of freedom*) to detect the abnormal events.

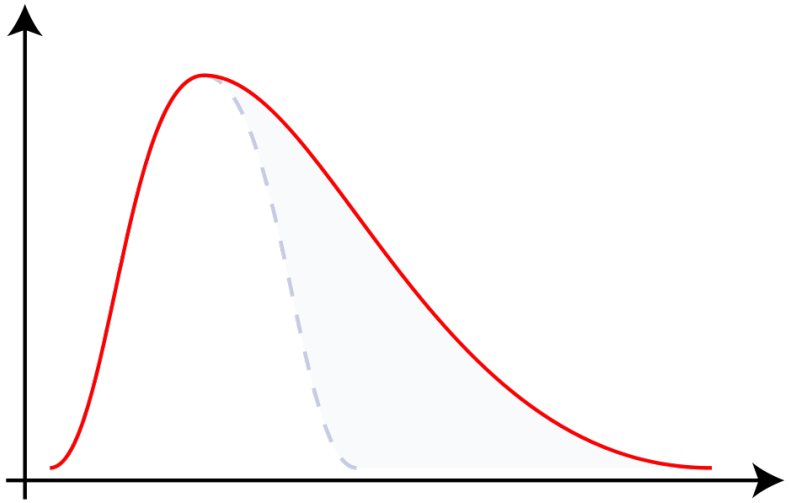
Normal and not normal



Normal

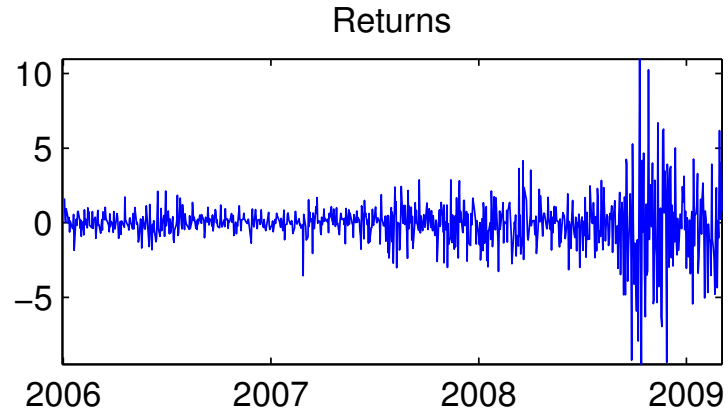


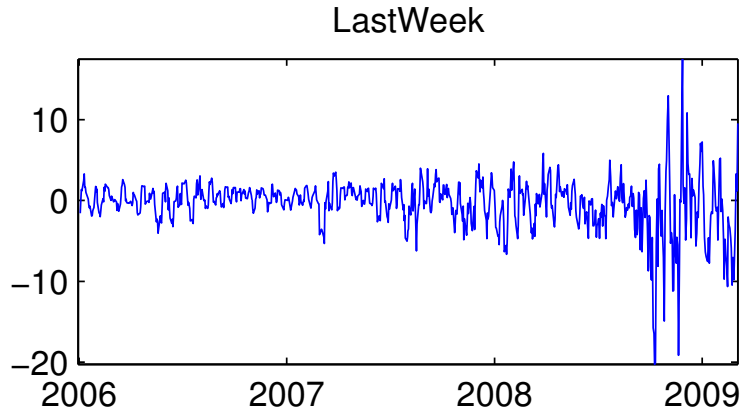
Negative Skew

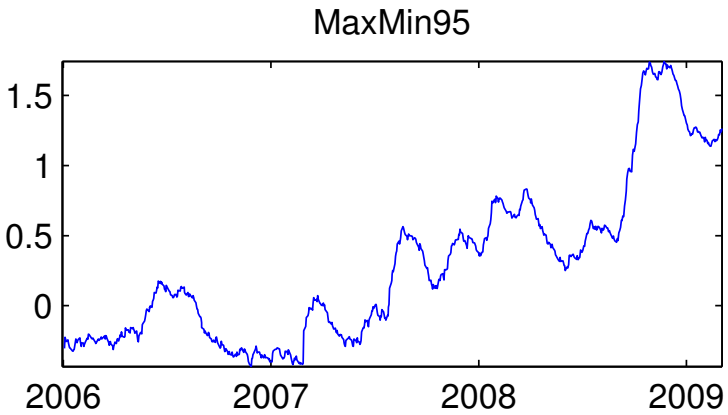


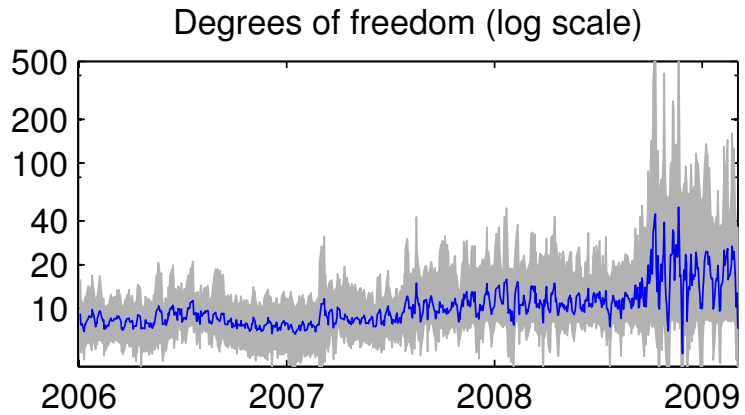
Positive Skew

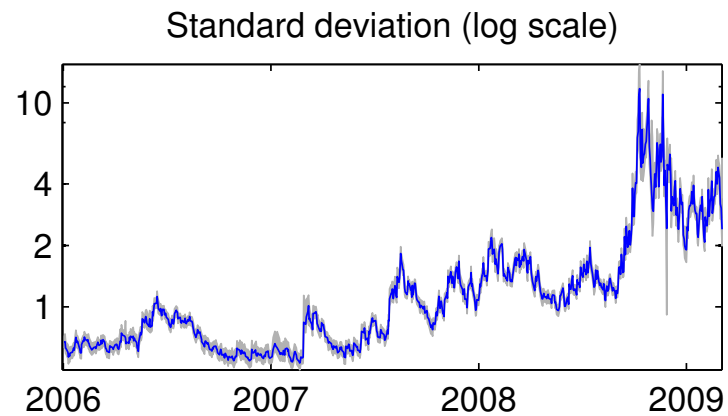
Detecting the financial crisis

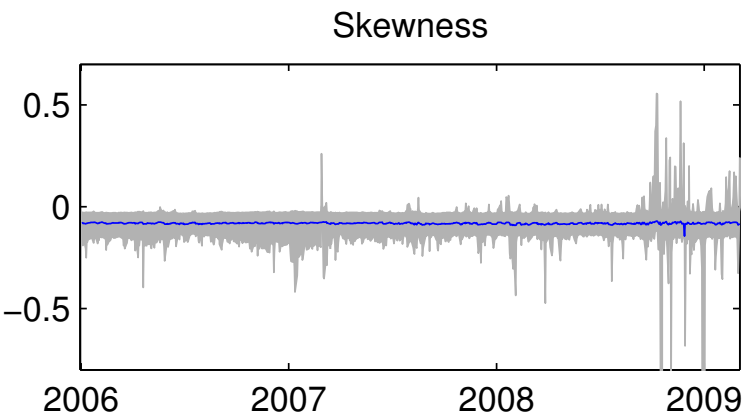








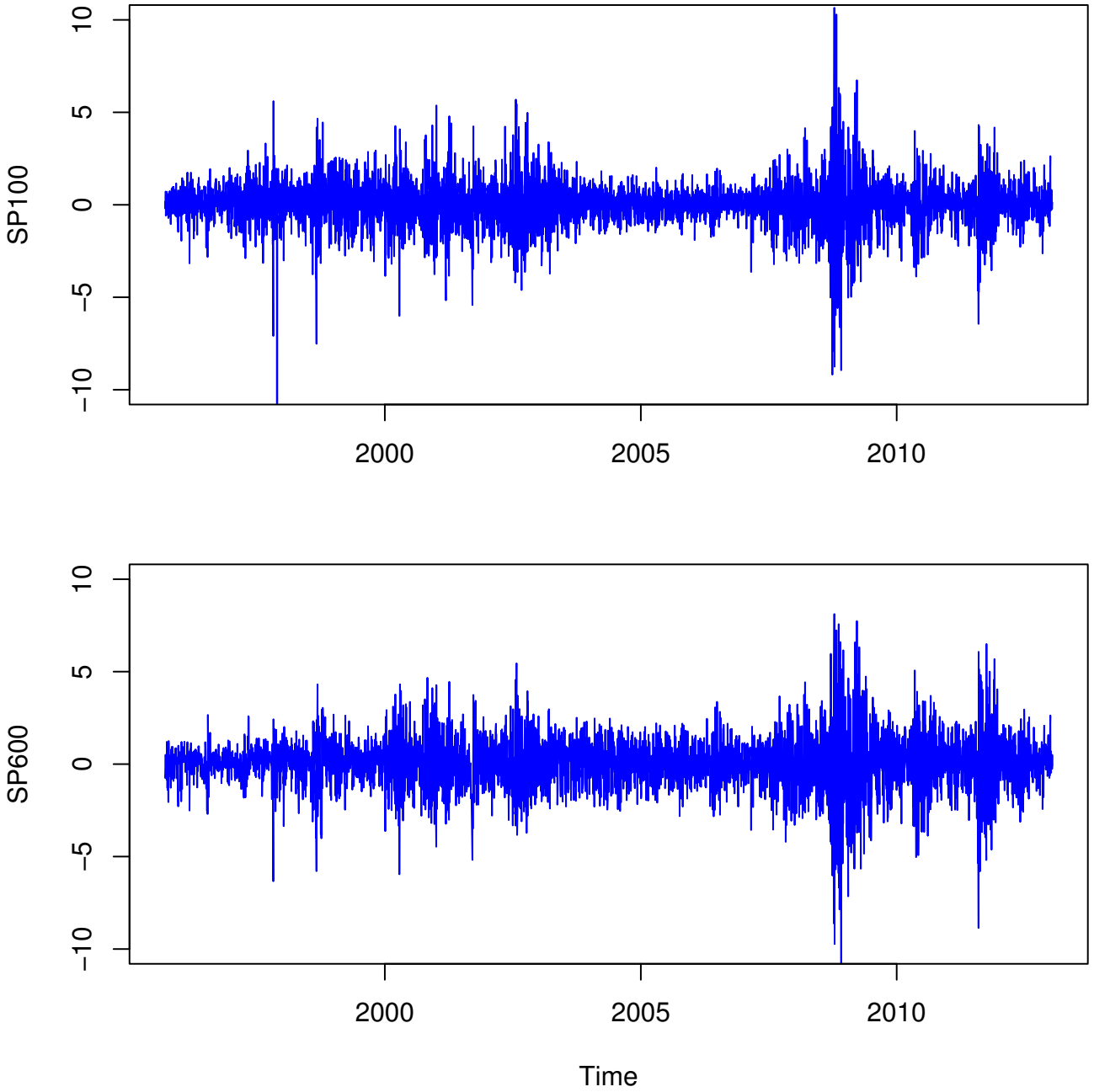




What if we insist using the normal model?

- The model will be misspecified.
- The conclusion based on that model can lead to a wrong decision.
- But people still do it anyway!
  - The normal model is simpler anyway.
  - We eventually do not know that we are wrong.
  - The computational tools are still feasible for everyone to use.
    - There was no ready-to-use computer software to use for this model.
    - The model takes a night to estimate with a cluster.

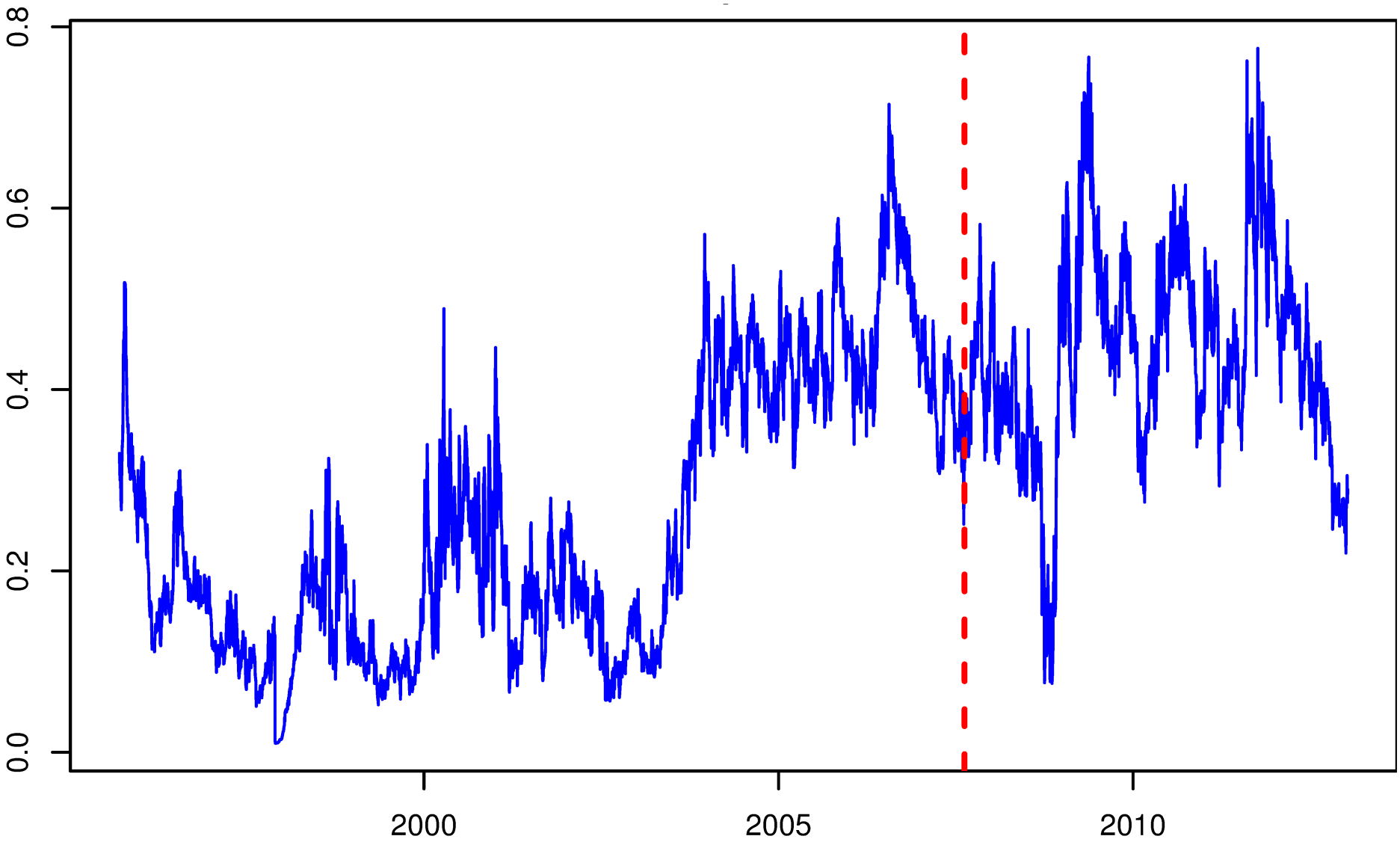
A more complicated situation



Our interests

- The S&P100 index includes the largest and most established companies in the U.S.
- The S&P600 index covers the small capitalization companies which present the possibility of greater capital appreciation, but at greater risk
- We are not only interested to detect the extreme events of a single stock, but also the co-movement of a two or more stocks.
    - What will happen to S&P100 if S&P600 collapses?
    - We call this as **tail-dependence**—the dependence only happens when extreme events happen.
    - What are the underlying factors that are connected to this dependence?

The dependence on the tail



Knowing the elephant
↛ The trend of statistical modeling

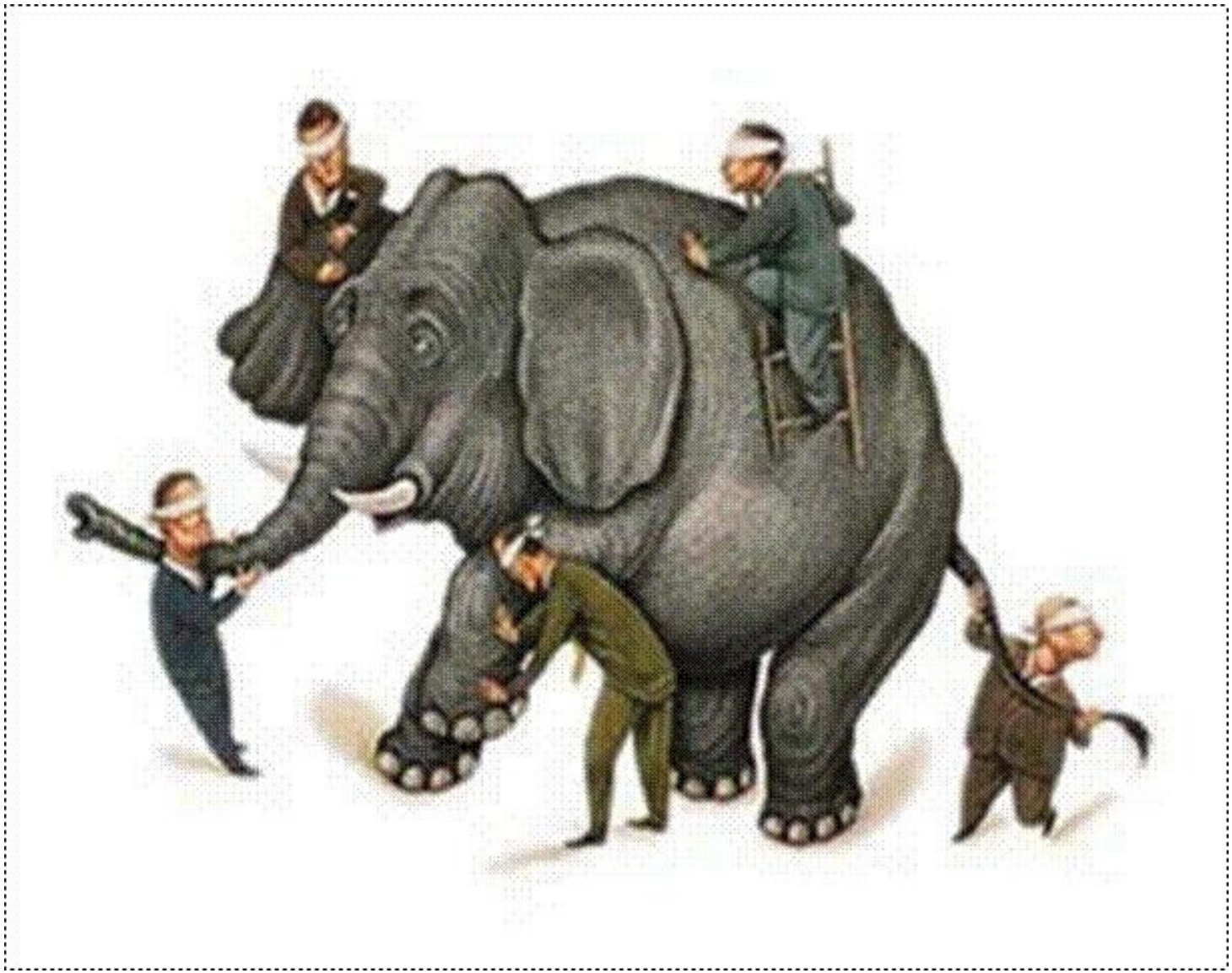
- In the 1950s, linear regression model was considered as very advanced which is now the standard course content for university students.
- The data are much more complicated nowadays we meet.
  - Numerical, categorical, brain image...
  - A few observations to millions by millions.
  - Very high-dimensional data are not rare anymore.

Knowing the elephant
↛ The common procedure statistical modeling

- Data collection
- Model estimation
- Model evaluation
- Model comparison
- Prediction (if needed)

Knowing the elephant

↛ Can we have a model that is big like an elephant?



Knowing the elephant

- Sophisticated models are essential for such situations.
- In principle, the complicated model should be able to capture more complicated data features.
- Estimating such model is not easy.
- There is huge space to explore.
  - The computer is still not fast enough.
  - Techniques like parallel computing should be used to speed up the computation.
  - Statistics with big data is the new challenge.

大数据基础

李丰

FENG.LI@CUFE.EDU.CN

中央财经大学

大数据时代最关心的几个问题

1.什么是大数据?
2.大数据能做什么?
3.哪里有大数据?
4.怎么分析大数据?

事实

- 48 %的高管认为，大数据是一个有用的工具，而另外23 %的人认为大数据将彻底改变企业的管理方式。
- 缺乏了解如何利用大数据直接阻碍了大数据在企业中的应用。

想像中的大数据


WHAT WE KNOW...
THE REST...
BIG DATA

大数据起源与发展

大：多，非常多，海量

数据：有价值，有指导作用

- Volume 数据大小
- Velocity 数据输入输出的速度
- Variety 多样性
- Variability 变化性
- Veracity 真实性
- Complexity 复杂性

- 1YB = 1,000 ZB
  - 截止 2011 年，没有任何人类制造的存储设备容量达到 1YB
- 1ZB = 1,000 EB
  - 据估计 2012 年全球所有存储数据为 2.7ZB
- 1EB = 1,000 PB
  - 1 克 DNA 中可以储存 360EB 的信息量。
- 1PB = 1,000 TB
  - 在 2009 年，阿凡达 3D 版使用了多于 1PB 来储存。
  - Google 的服务器场的估计容量大约是 5.625PB(截止 2004 年) 。
  - 国会图书馆（美国）藏品数量的估计近似值大约是 10PB （包括非书籍资料，截止 2005 年）
  - 在 2012 年的 8 日 , Facebook 共使用了约 100PB 。
  - 全世界印刷材料的内容总量大约是 300PB 。
- 1TB = 1,000 GB
  - 绝大多数笔记本硬盘的容量

- 据 IBM 称， 90 %的组织收集的数据为商业数据。
- 社会化媒体在这份名单上排名最低，仅为 39 %。
- 全世界 90% 的数据在过去两年被创造。

大数据起源与发展

起源于搜索引擎
（ Google, 百度 ）

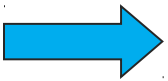
社交网络，电商等互联网领域丰富发展（ Facebook, QQ, Alibaba ）

全民大数据

大数据应用经典案例

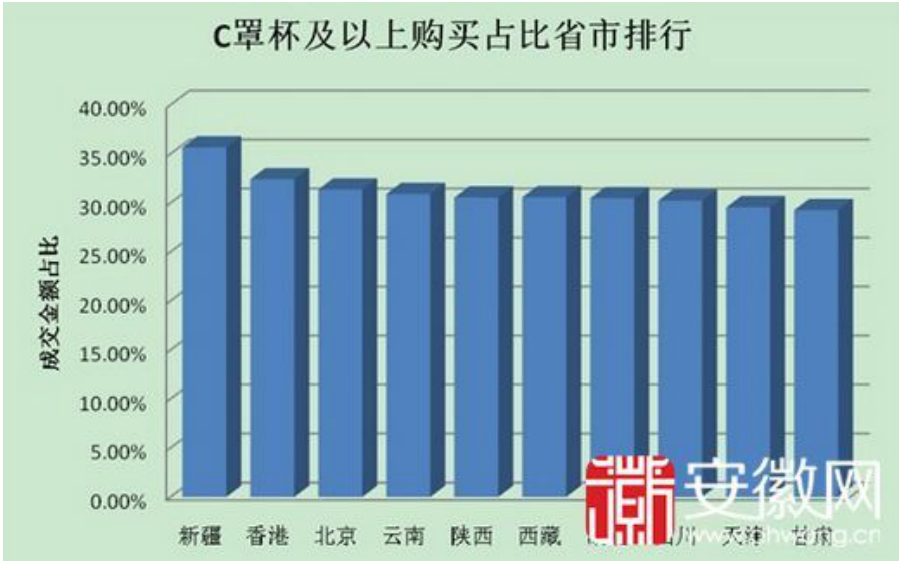
经典教科书案例 --- 奶爸与啤酒

Prada,RFID, 大数据分析



大数据应用经典案例

淘宝数据魔方



类似产品：京东罗盘，微数据等

大数据应用经典案例

QQ 圈子把前女友推荐给未婚妻

大数据解决方案：医药销售预测

- 传统数据：各种药品门店销售数据
- 困扰：
  - 如何精准的控制库存
  - 如何预测接下来五天的进药量
  - 门店销售人员不懂大数据，甚至不会操作 Excel
- 解决方案：
  - 结合外部数据：天气数据
  - 数据上云，中央解决
  - 门店销售人员配备 APP

利用数据科学工具获取一手数据

李丰

统计与数学学院
feng.li@cufe.edu.cn

学习“数据科学工具”的必要性

- “我有A股企业全部PDF格式的财务报表，如何从中迅速找到各个企业的净利润项?”
- “什么数据能够支持我研究散户的情绪是如何影响大盘的？我从哪里能下载到？”
- “研究网络借贷平台融资的数据如何得到？”

……

目标

- 培养在大数据时代利用现代数据科学工具独立获取**一手数据**的能力。

- 训练利用数据分析**工具链**发掘复杂数据价值的数据科学思维模式。

数据科学工具学什么内容

一：数据科学核心价值	五：数据可视化
二：现代数据生产环境与工具	六：海量数据分布式计算基础
三：创建自己的大数据仓库	七：大数据价值实现
四：非结构化数据处理基础	八：数据科学应用案例

科学素养与一体化的知识结构

- 数据科学不断改变业界、学者、学生的思维模式。加强学生数据科学工具使用的培养，既是对大数据时代的融入，也是未来的方向。
- **现代大数据背景下**强调以实际应用为导向的知识思维和学习模式，避免初学者学习目的盲目的问题，又有效地平衡和数据科学内容繁复与学习精力有限的问题。
- 在方法论上，应体现利用数据科学知识与专业知识结合，培养跨界融合的创新精神。
- 在科学实践中，教师应积极引入现代交互式教学模式，可以不断校正学习模式，提高发现、利用一手数据探究问题本质的能力。

感谢各位聆听！