Do Google Trends data contain more predictability than price returns?

D. Challet[1,2] A. Bel Hadj Ayed [1]

[1] École Centrale Paris [2] Encelade Capital SA

25th November 2014

Google Hedge Fund

In "Googled: The End of the World As We Know It", Ken Auletta

Sergey Brin: "We should run a hedge fund."

Eric Schmidt: "Sergey, among your many ideas, this is the worst"

Sergey Brin: "No, we can do it because we have so much information."

Eric Schmidt: "[...] legal complications [...] NO!"

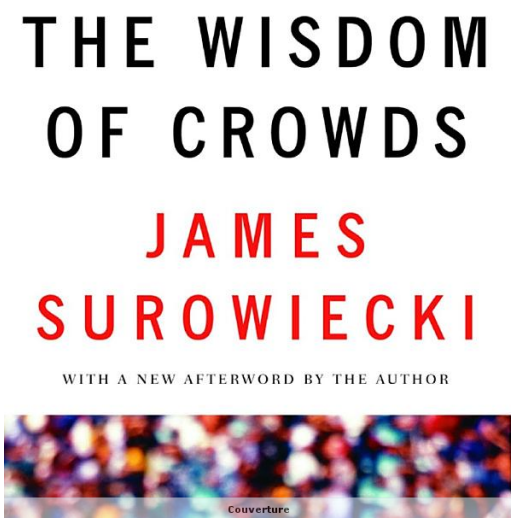
Prevailing wisdom about the crowds

No one in this world, so far as I know, has ever lost money by underestimating the intelligence of the great masses of the plain people.

The crowd [...] average intelligence is very low; it is inflammatory, vicious, idiotic, almost simian

H. L. Mencken (essayist, satirist)

Wisdom of the crowds



Why the Many Are Smarter Than the Few

and How Collective Wisdom Shapes

Business, Economies, Societies

and Nations

Galton 1906

Guess the weight of a fat ox, after slaughter and dressing

787 tickets

Degrees of the length of Array 0°—100°	Estimates in lbs.	Centiles: Observed deviates from 1207 lbs.	Centiles: Normal p.e =37	Excess of Observed over Normal
5	1074	−133	−90	+43
10	1109	− 98	−70	+28
15	1126	− 81	−57	+24
20	1148	− 59	−46	+13
q_1 25	1162	− 45	−37	+ 8
30	1174	− 33	−29	+ 4
35	1181	− 26	−21	+ 5
40	1188	− 19	−14	+ 5
45	1197	− 10	− 7	+ 3
m 50	1207	0	0	0
55	1214	+ 7	+ 7	0
60	1219	+ 12	+14	− 2
65	1225	+ 18	+21	− 3
70	1230	+ 23	+29	− 6
q_3 75	1236	+ 29	+37	− 8
80	1243	+ 36	+46	−10
85	1254	+ 47	+57	−10
90	1267	+ 52	+70	−18
95	1293	+ 86	+90	− 4

q_1, q_3, the first and third quartiles, stand at 25° and 75° respectively.
m, the median or middlemost value, stands at 50°.
The dressed weight proved to be 1198 lbs.

Mathematical wisdom

median/average of estimates better than best estimate

Wisdom of the crowds: success and dangers

Good

- one-shot
- simultaneous choices
- no communication

Bad

- asynchronous choices + communication
- repeated game

Prediction markets?

More than 787 votes?

Active users

Google ..... 1.1×10^9

Facebook ..... 1.1×10^9

Google+ ..... 0.5×10^9

Twitter ..... 0.2×10^9

Publicly available data

Google . . . . . aggregate

Facebook . . . . . ?

Google+ . . . . . posts

Twitter . . . . . posts (?)

What to do with this data

Nowcasting

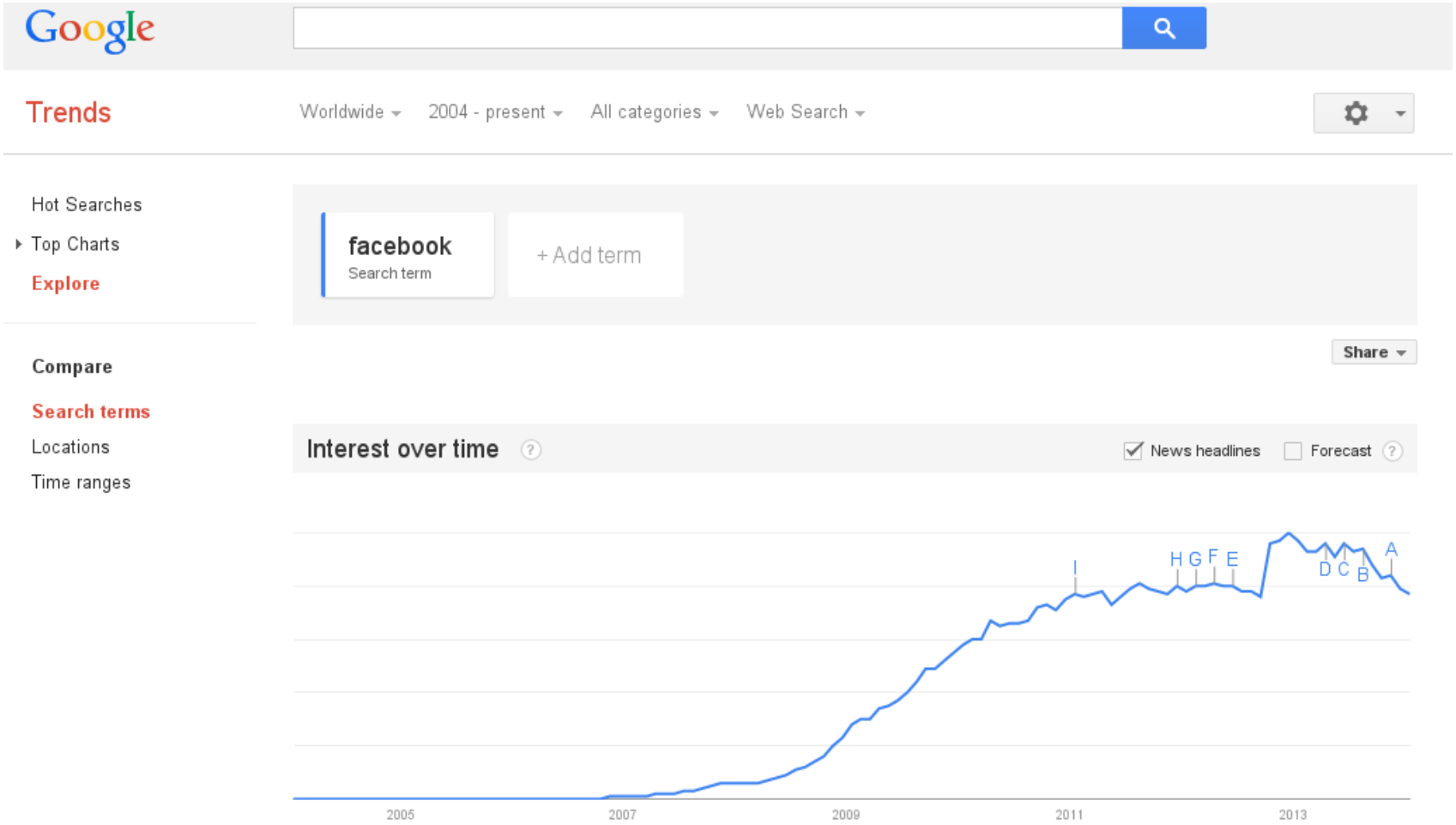
- Users create data
- Activity correlated with some quantity
- Social data $\longleftrightarrow$ macroscopic variable

Choi and Variant (2009): Predicting the Present with Google Trends

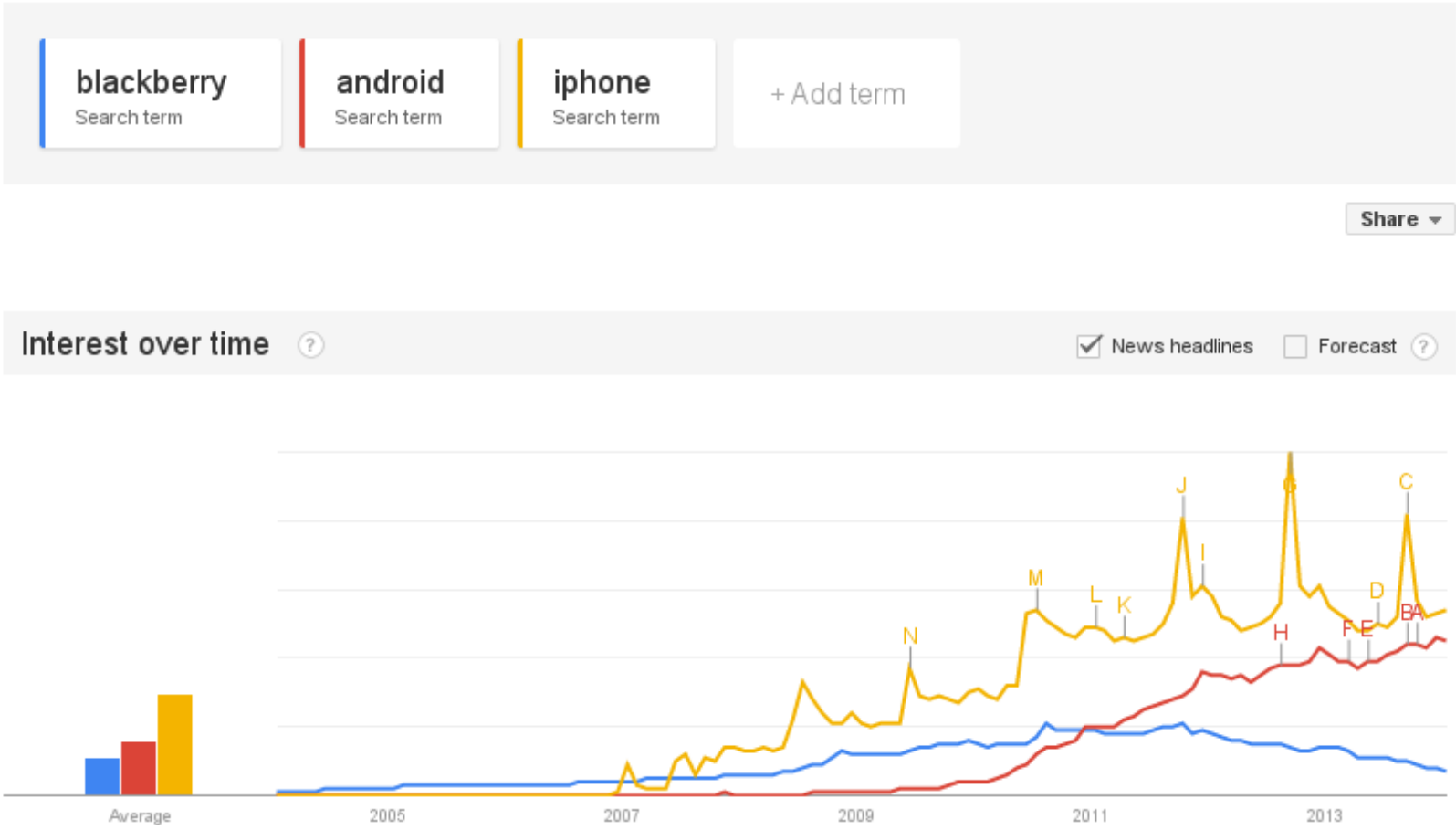
Forecasting

- Users create social data
- (some) Users trade
- Social data $\longleftrightarrow$ future price changes

Google Trends: Search Volume Index



GT: example


blackberry
Search term
android
Search term
iphone
Search term
+ Add term
Share
Interest over time
News headlines
Forecast
Average
2005
2007
2009
2011
2013

Nice features of GT data

The good

- Many users
- Aggregate
- Clean
- Download csv file
- Any keyword

The bad

- Anonymous
- Irrelevant information
- Coarse
- ?
- Too much choice

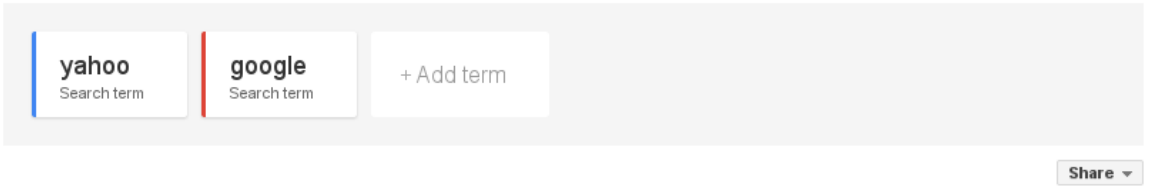
GT and predictability: claims

Keywords: ticker, company names

- Bordino et al. 2011,
  increase in SVI $\rightarrow$ *increase of traded volume*
- Da et al 2013, [2004-2008]
  increase in SVI $\rightarrow$ *higher stock prices in the next 2 weeks*
- Joseph et al 2011 [2005-2008],
  increase in SVI $\rightarrow$ *higher stock prices in the next week*
- Takeda et al. 2013 [2008-2011]:
  weak for future returns, strong for future volume
- Kristoufek 2013 [2004-2013]:
  portfolio weight $\sim SVI^{-\alpha}$
- Preis et al 2013 [2004-2011]:
  fancy keywords;
  relative increase in SVI $\rightarrow$ *lower DJIA the next week*

$\uparrow$meaning of GT invariant among stocks AND time?

Meaning of GT invariant among stocks



Meaning of GT invariant across time

- Forecasting: Da et al. (2013)

	Lagged 1 Week				
	log(SVI)	log(turnover)	Absolute Abn Ret	log(1 + Chunky News)	R^2
log(SVI)	0.5646***	−0.0022***	0.0489***	−0.0027***	56.47%
	0.01	0.01	0.01	0.01	0.01
log(turnover)	0.0532**	0.4467***	0.5197***	−0.0298***	38.82%
	0.05	0.01	0.01	0.01	0.01
Absolute Abn Ret	0.0046***	0.0015***	0.0418***	−0.0011***	3.55%
	0.01	0.01	0.01	0.01	0.06
log(1+Chunky News)	0.0683**	0.0270***	0.2071**	0.0197***	3.19%
	0.02	0.01	0.05	0.01	0.01

- Google Trends: demand of information

$$return_t \sim F\left[\frac{Demand_t - Supply_t}{\lambda_t}\right]$$

- $Supply_t$ and λ_t !!!

A practitioner point of view

1. Trading strategies
2. Backtest period
3. Assets
4. Keywords
5. Download GT data
6. Timescale of returns
7. Parameters
8. Input GT data only,
9. Input past returns only
10. Input both
11. Compare.

1. Trading strategies

- ~~Linear methods~~
- Conditional predictability
- Ensemble learning methods (Support Vector Machines, trees, etc)

2+3 Backtest period, assets

2. Backtest period
   - ~~All previous papers: Whole period~~
   - Sliding in/out-of-sample periods

3. Choice of assets
   - Index components: S&P100

Recipe for disaster:

1. Think of finance-related keywords

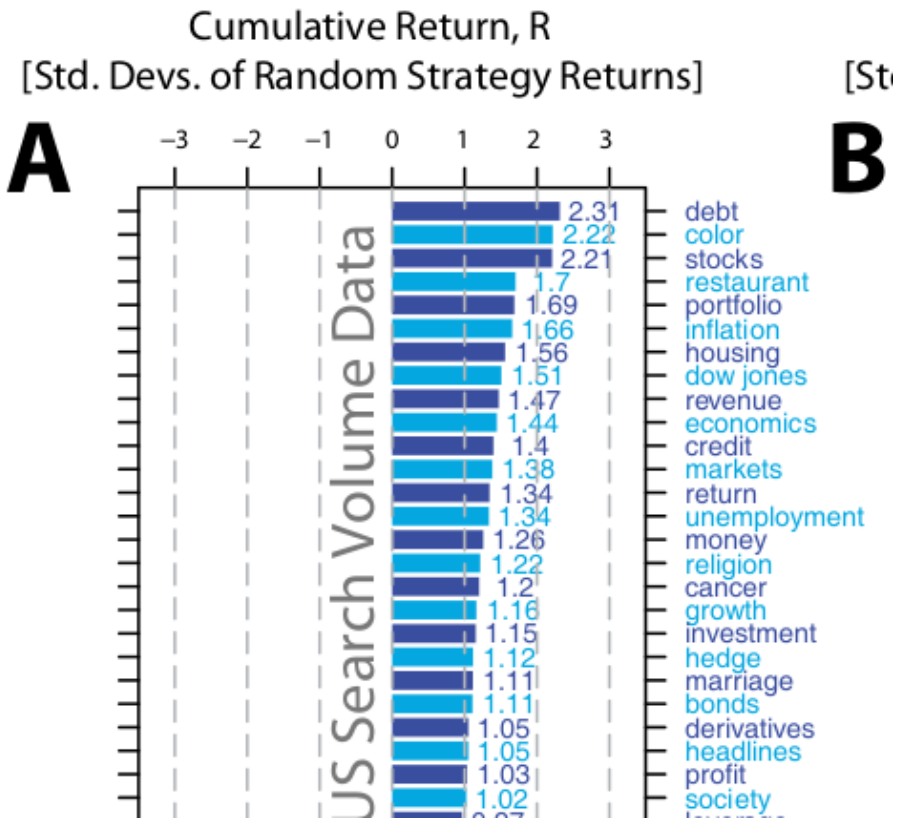
   finance, debt, CDS, bonds, crisis

2. Use Google Sets:

   finance $\rightarrow$ marketing, real estate, insurance, accounting,
   debt consolidation, investing,
   [...]

4. Keywords: example

Preis *et al.* (2013): moving averages (k weeks)

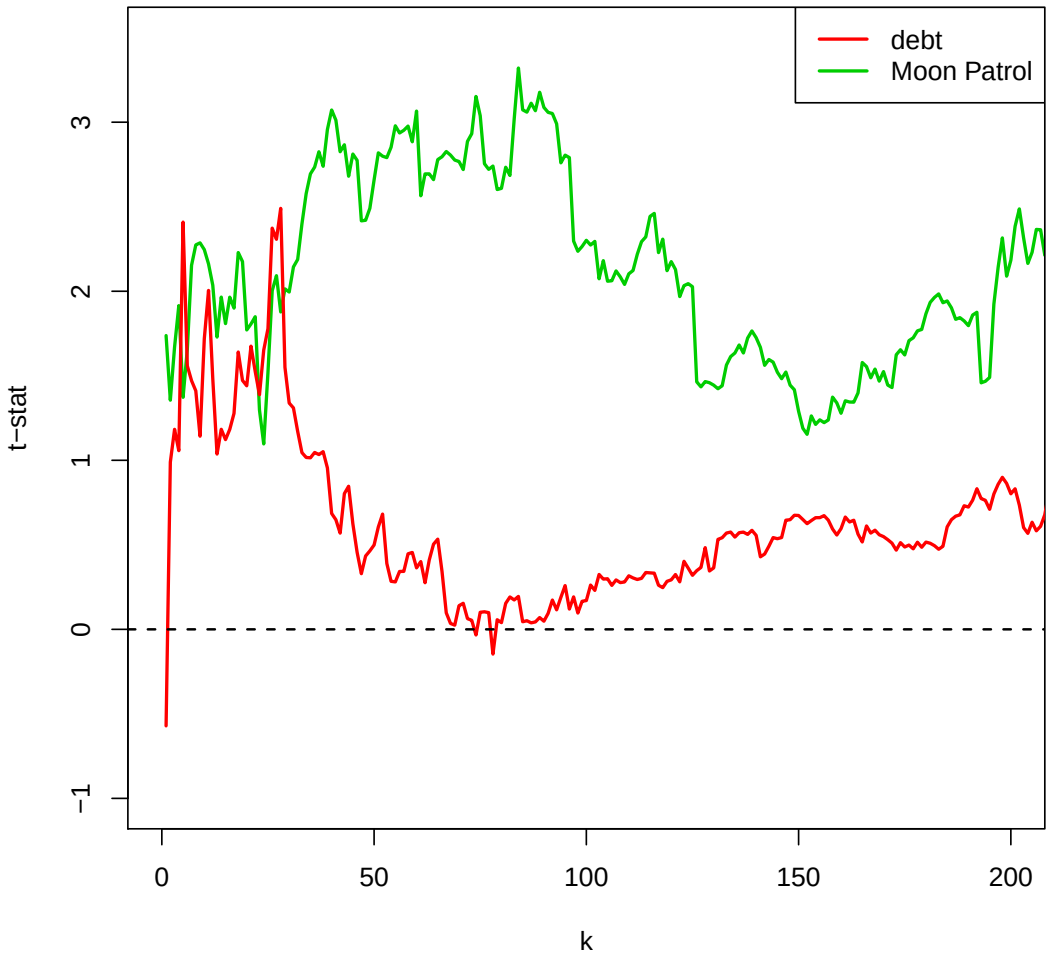


4. Keywords: null hypothesis?

1. 100 classic cars
2. 100 classic arcade video games
3. 200 classic illnesses/ailments

keyword	t-stat	keyword	t-stat	keyword	t-stat
multiple sclerosis	-2.1	Chevrolet Impala	-1.9	Moon Buggy	-2.1
muscle cramps	-1.9	Triumph 2000	-1.9	Bubbles	-2.0
premenstrual syndrome	-1.8	Jaguar E-type	-1.7	Rampage	-1.7
alopecia	2.2	Iso Grifo	1.7	Street Fighter	2.3
gout	2.2	Alfa Romeo Spider	1.7	Crystal Castles	2.4
bone cancer	2.4	Shelby GT 500	2.4	Moon Patrol	2.7

4. Keywords



IN SAMPLE

4. Which keywords?

KISS:

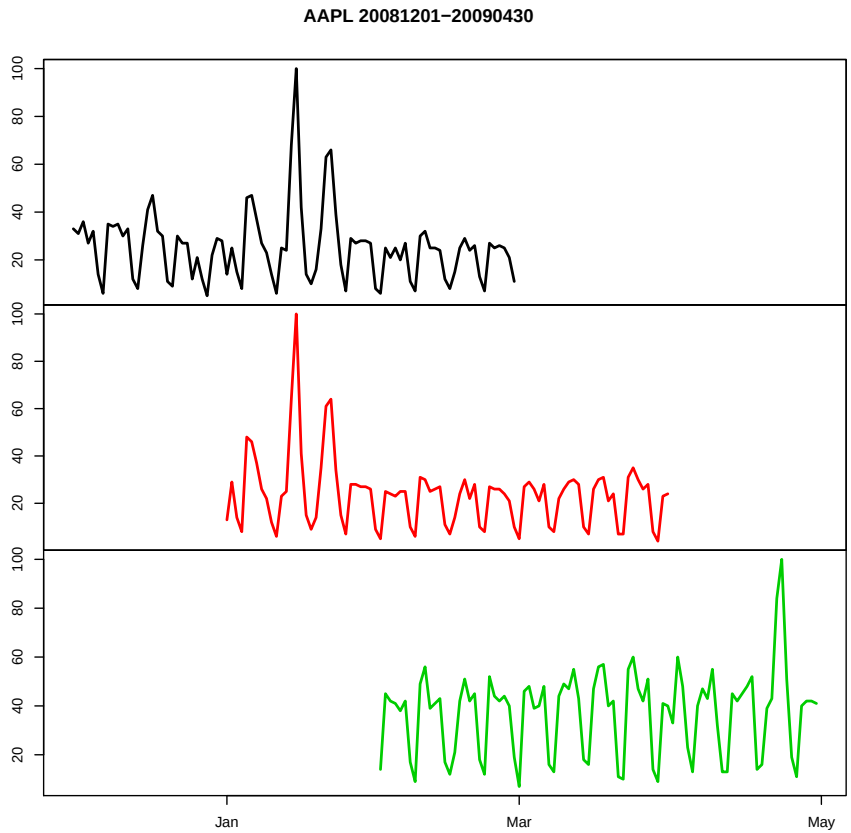
- Symbols
- Company names
- Key products

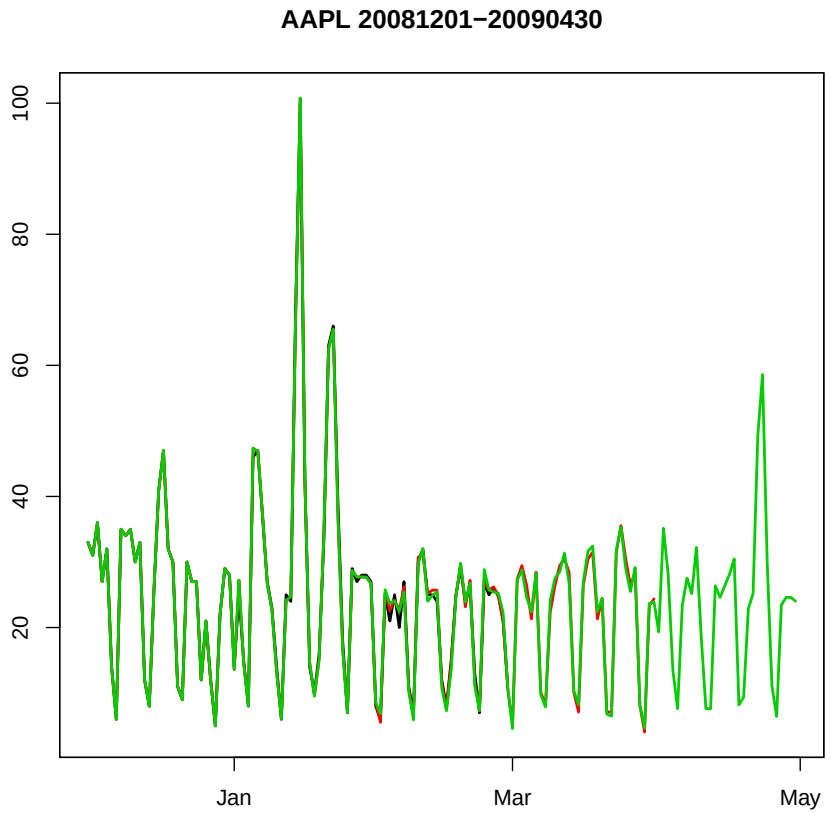
5. GT data

1. Weekly
2. Starts in 2004
3. Data not available before 2008-08

4. File format change in 2012-01
   - before
     ```
     Nov 27 2005, 1.14, 5%
     Dec 4 2005,  1.00, 5%
     ```
   - after
     ```
     2005-11-27 - 2005-12-03,31
     2005-12-04 - 2005-12-10,28
     ```

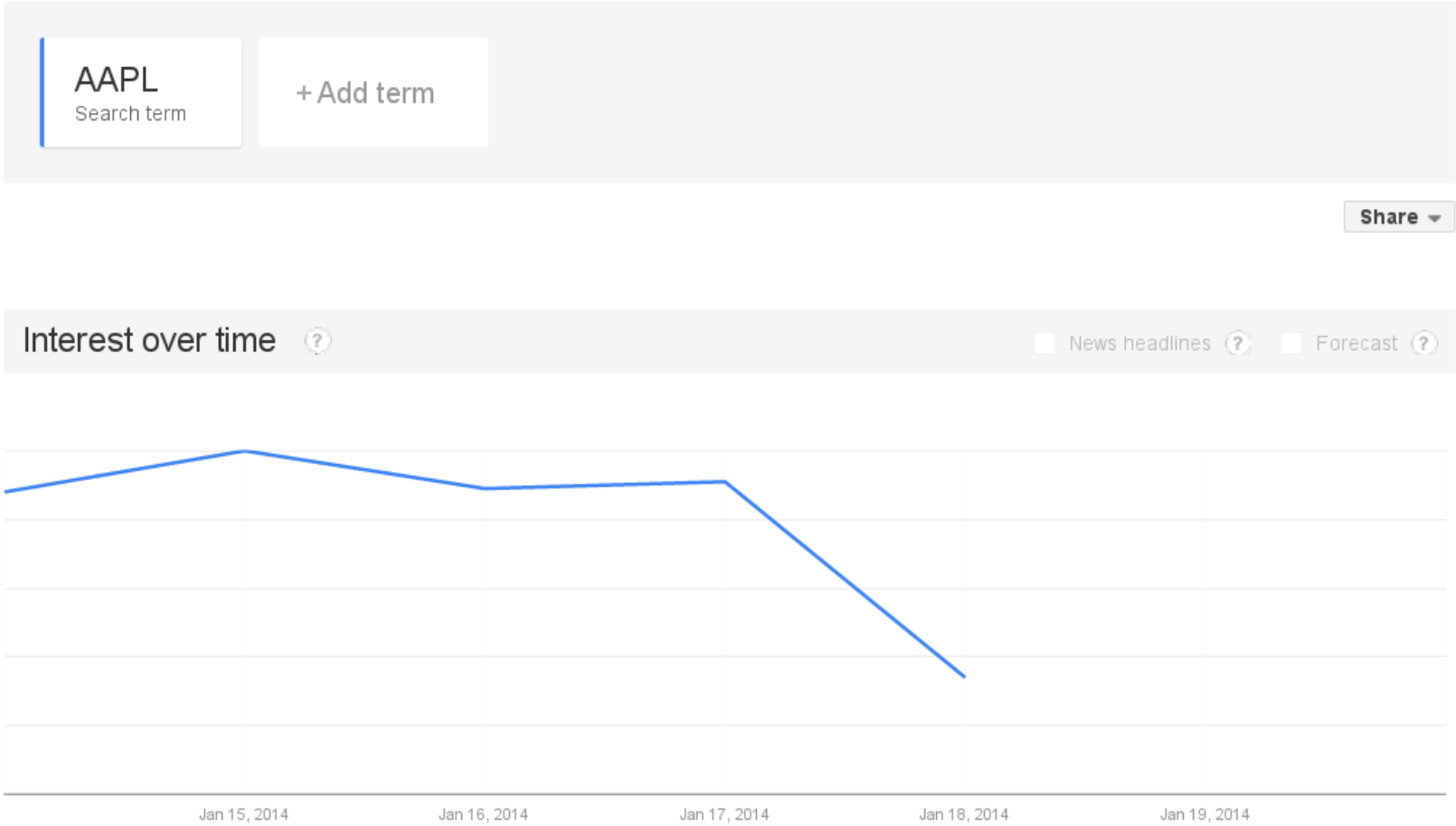
5. GT daily data





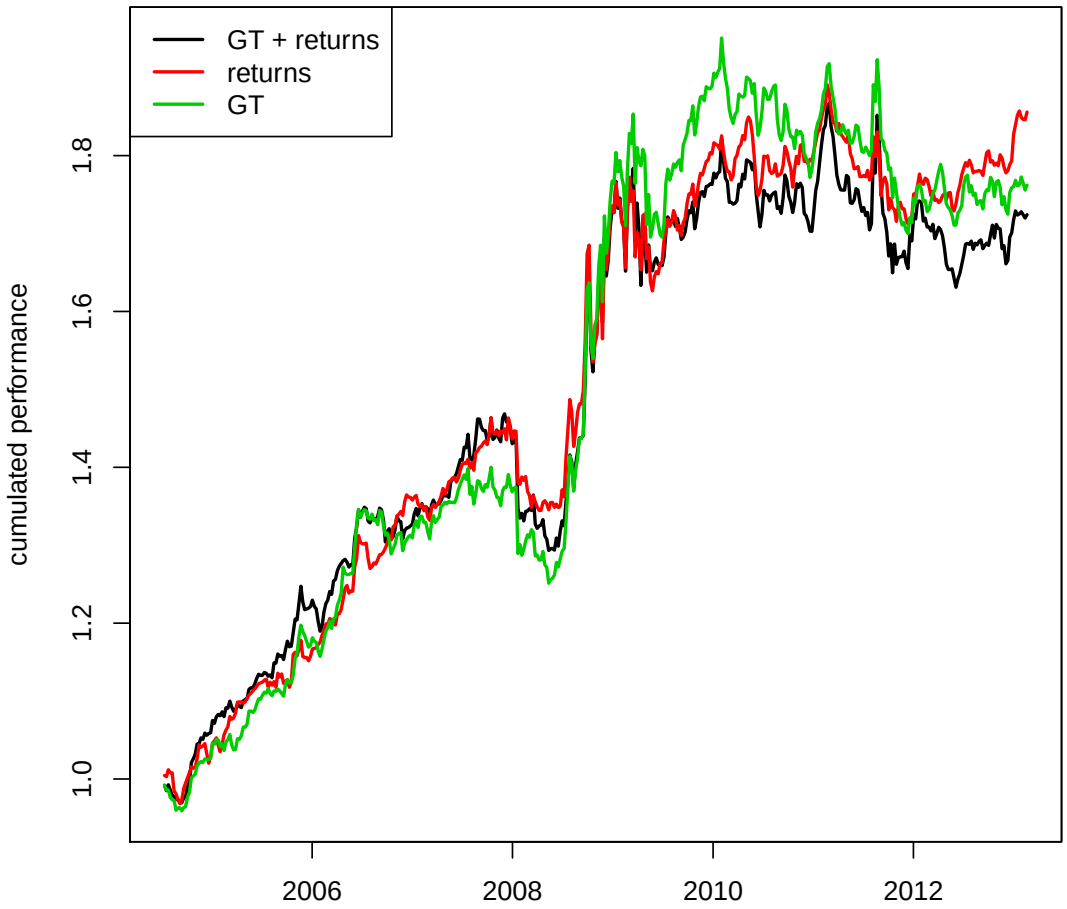
5. GT daily data: delay

(downloaded 2014-01-20, 09:02:00 UTC)



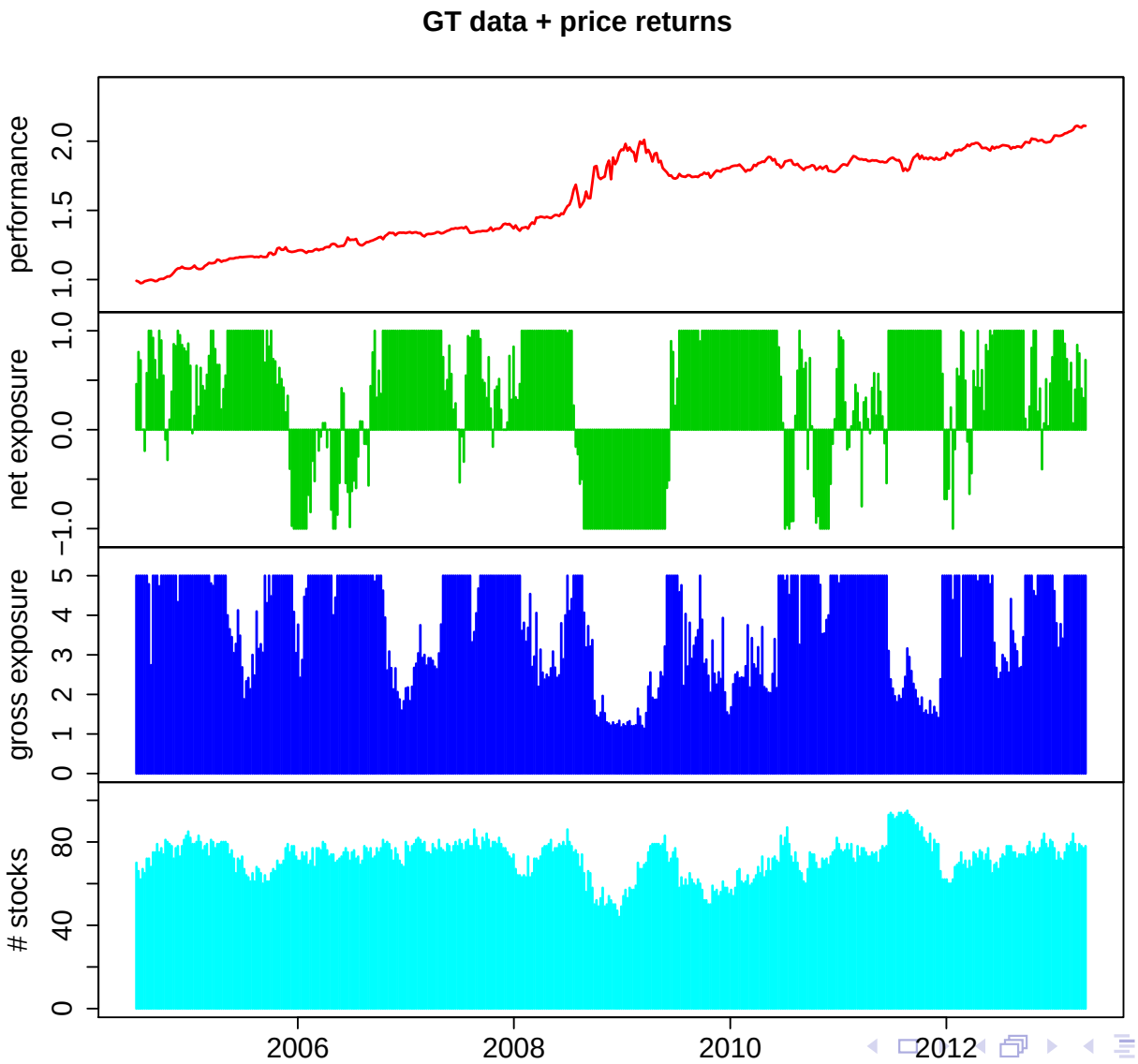
Backtest: Support Vector Machines

- 6 months in-sample
- 1 week out-of-sample
- Predictors: GT or returns, or both

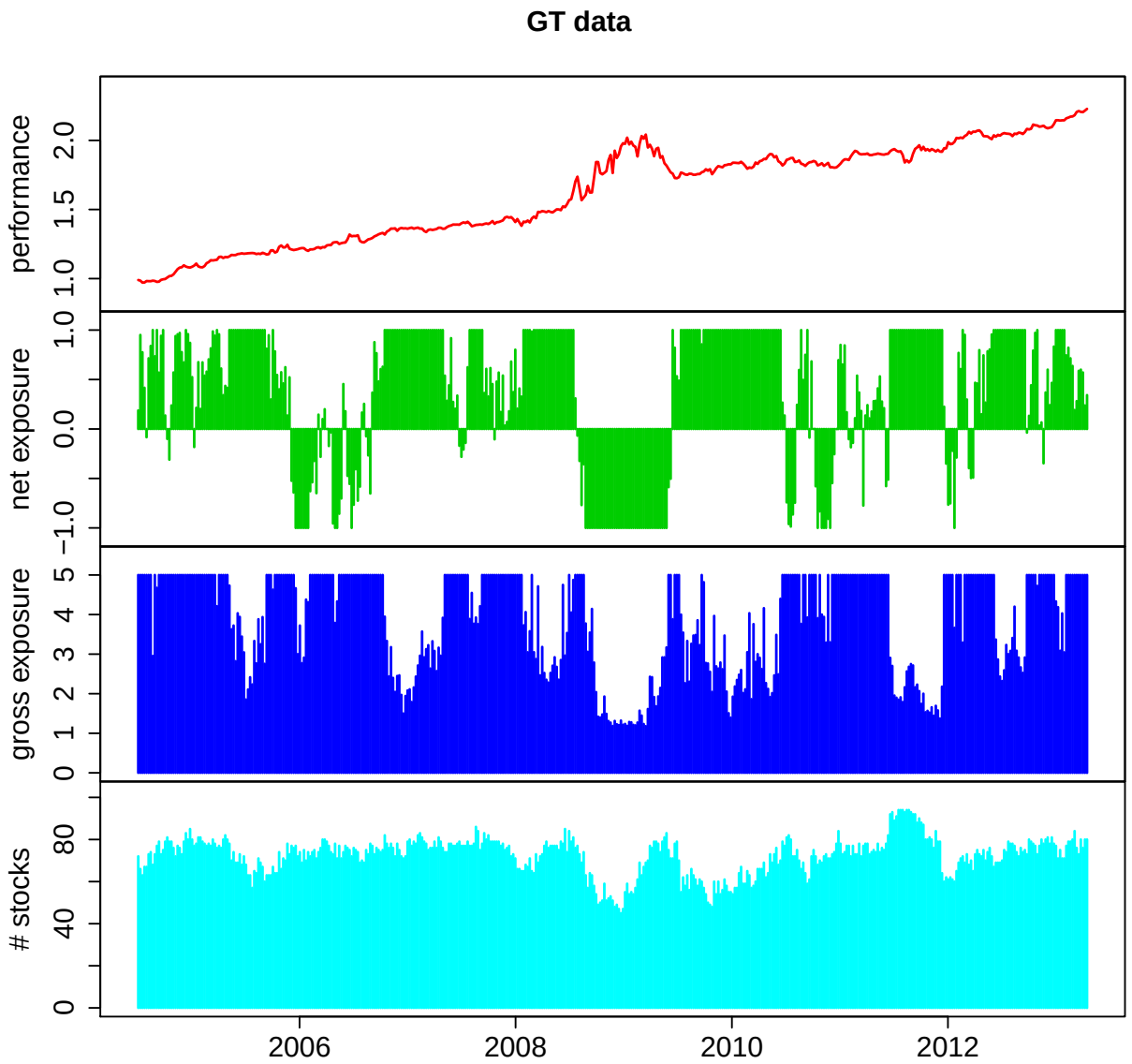


Backtest: GT + returns

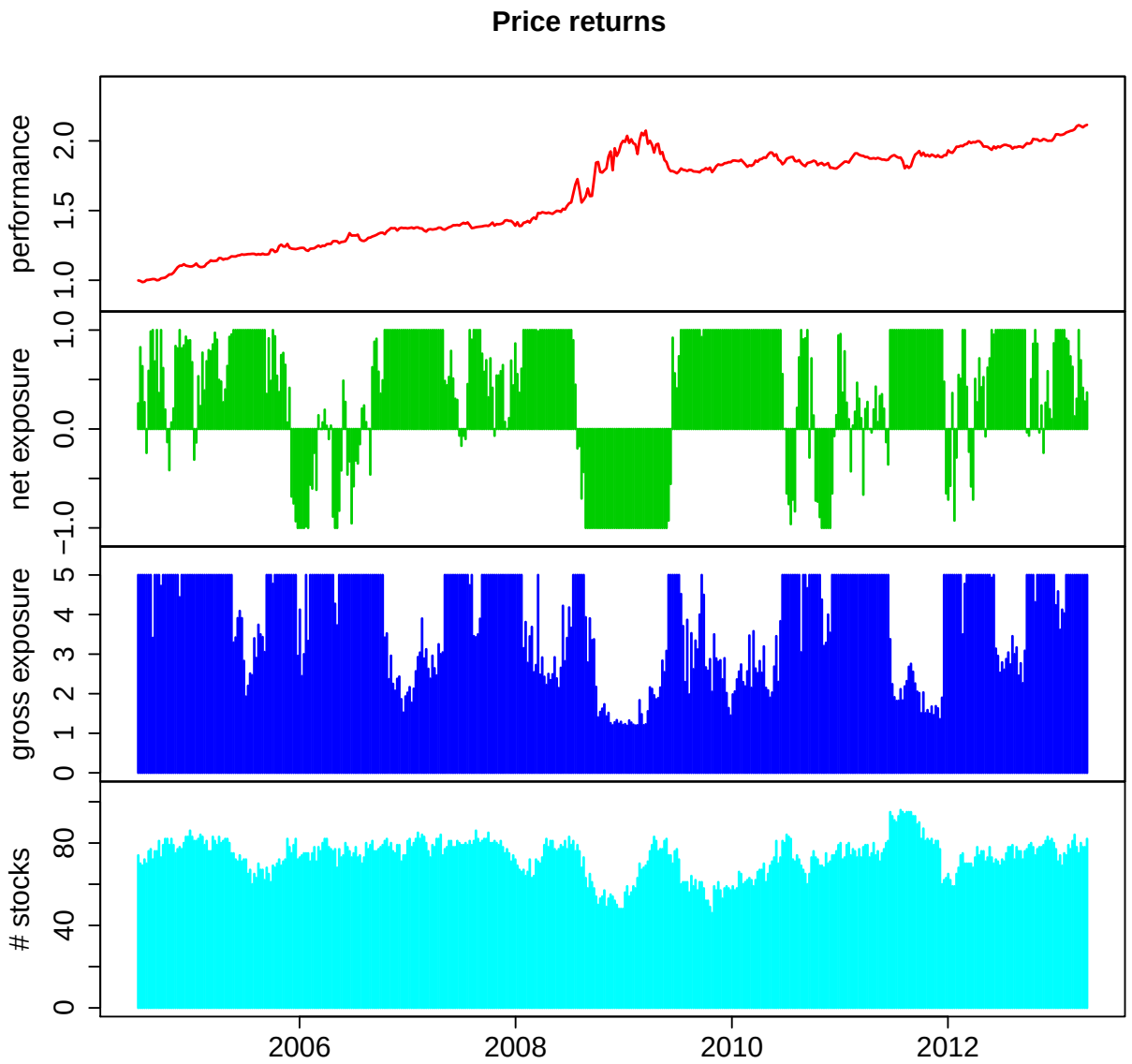
Six-month in-sample, ensemble learning+a few tricks



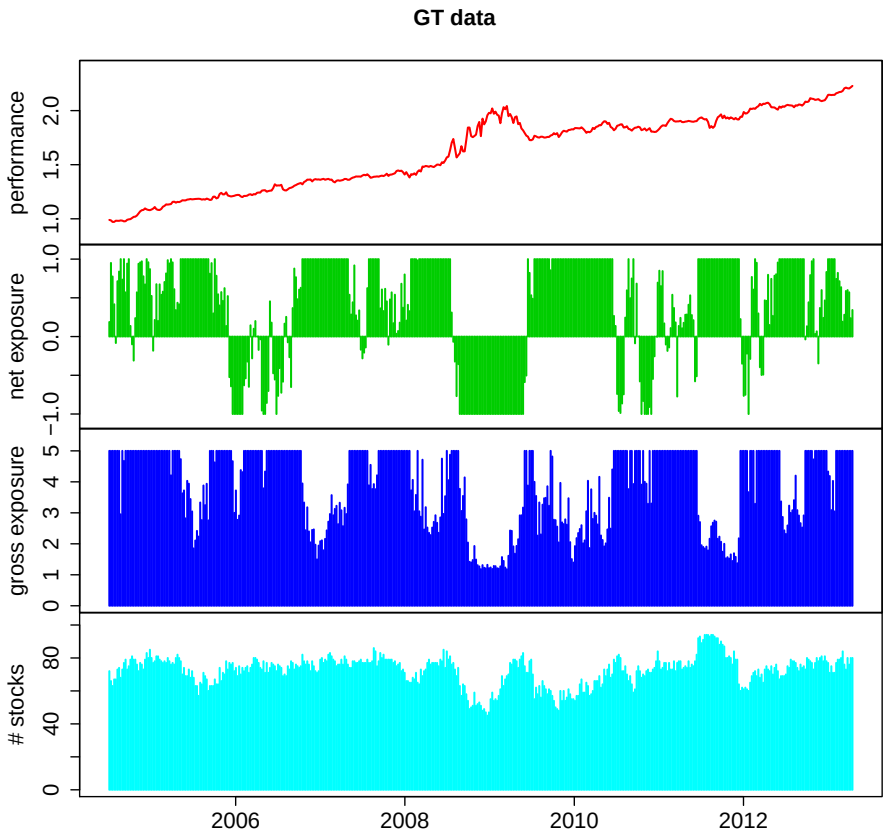
Prediction: GT data only

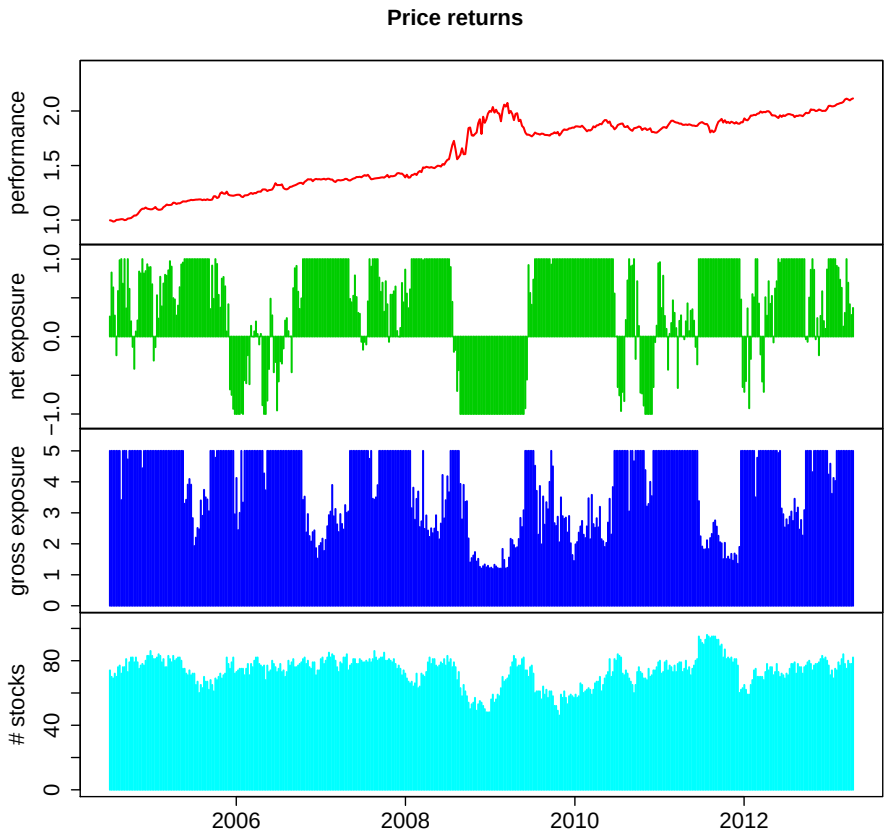


Prediction: returns only



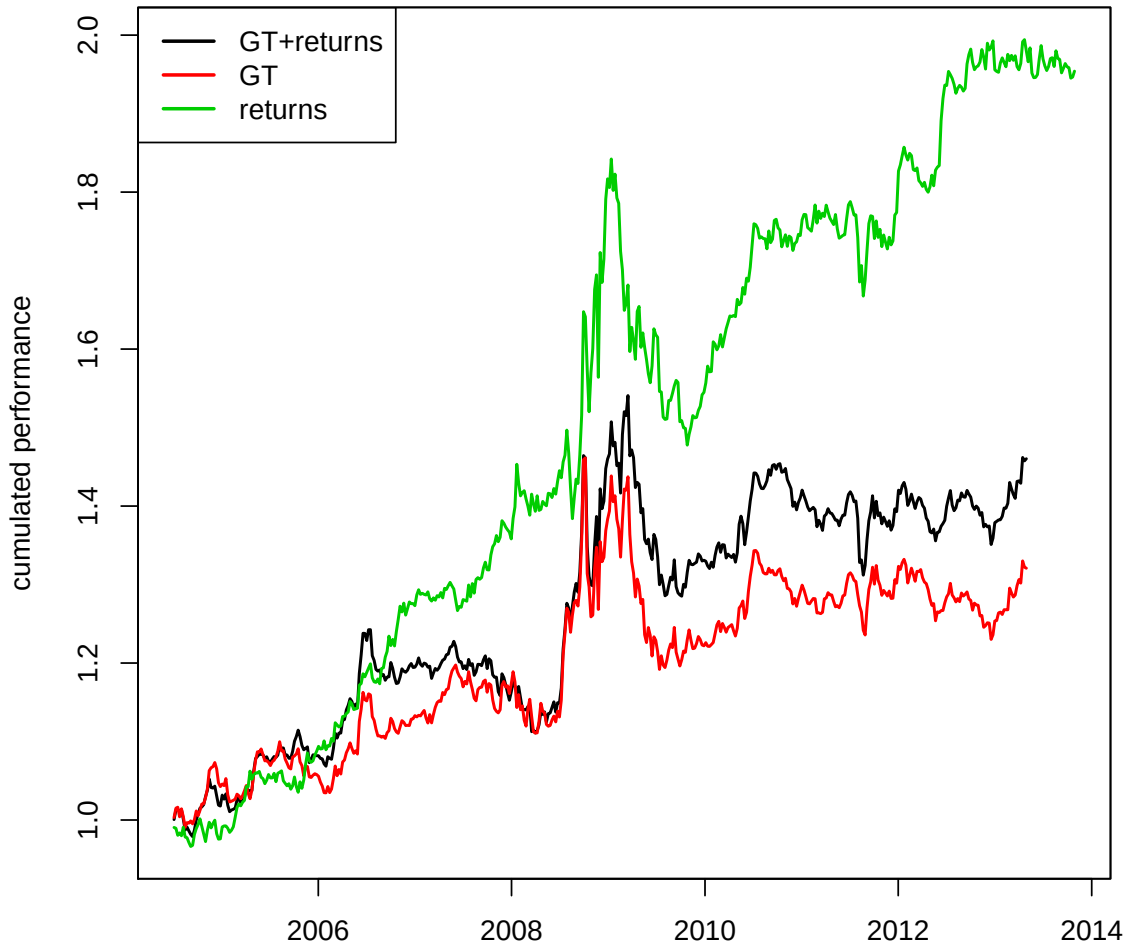
Prediction: comparison





Bug?

same backtest system, same parameters, binary inputs



Therefore

GT data $\equiv$ price returns

- Large price returns $\longrightarrow$ large GT change
- Noisy returns (2%)$\equiv$ noisy GT data (5%)
- Aggregate quantities

GT! Why, oh why?

1. Weekly data, lagged daily data
2. Keywords: ambiguous
3. ALL users, not only traders
4. Meaning of *more searches* varies
    1. pre-2008: more *SVI*, higher price
    2. 2008: less *SVI*, smaller price
    3. post-2008 ?
5. Google Trends: demand for information

Next?

We want

- Big data without the mess
- Precise queries
- Fine time resolution (< 1 day)

Wikimedia hourly dumps (20140120 at 12:09 UTC)

- pagecounts-20140120-070000.gz, size 93M
- pagecounts-20140120-080000.gz, size 94M
- pagecounts-20140120-090000.gz, size 98M
- pagecounts-20140120-100000.gz, size 104M
- pagecounts-20140120-110000.gz, size 167M