# Can internet search queries help to predict stock market volatility?

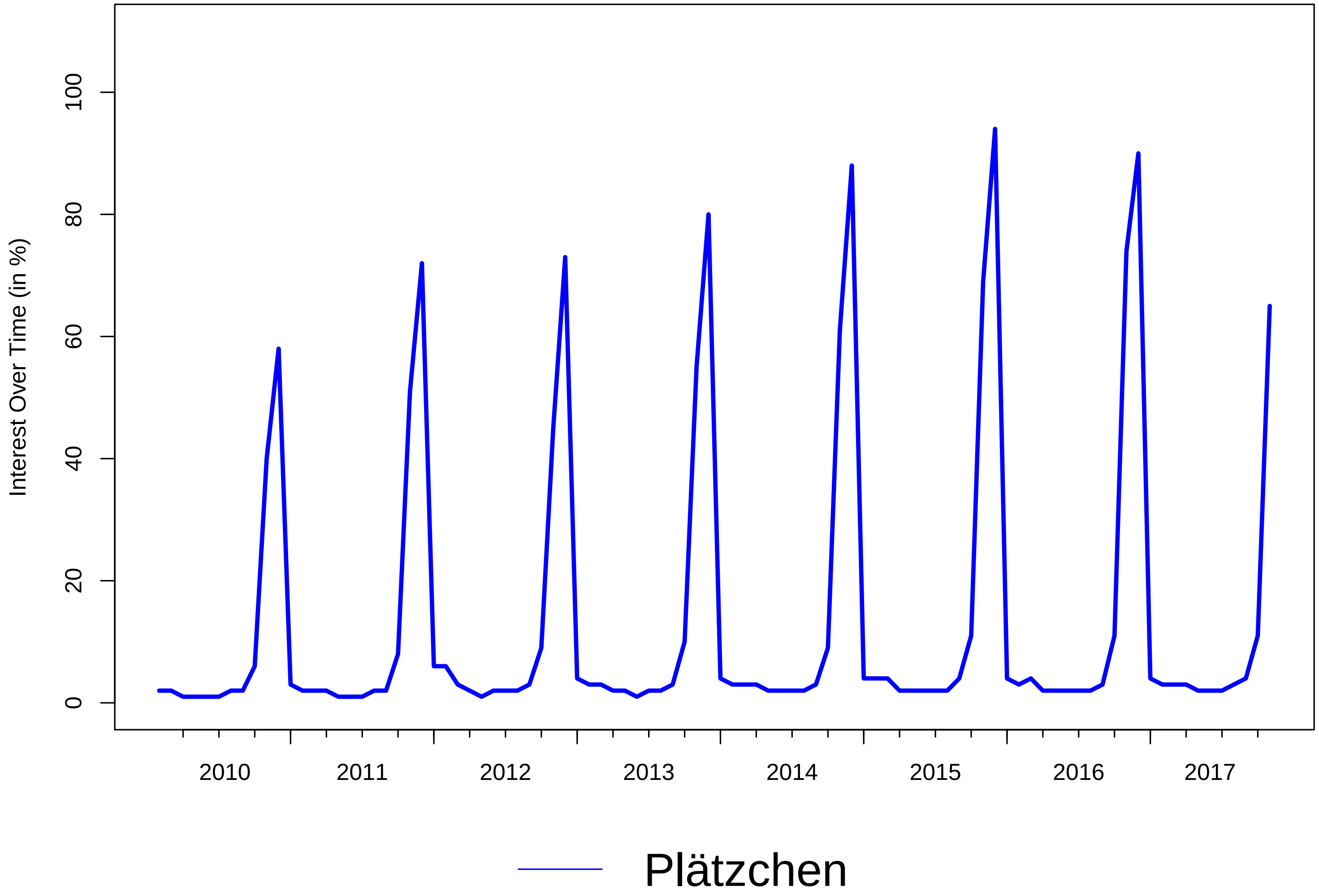
Thomas Dimpfl and Stephan Jank
Eberhard Karls Universität Tübingen

BFS Society Vortragsreihe
Tübingen, 4 December 2017

# Motivation

# Internet searches are a measure for individuals' interests

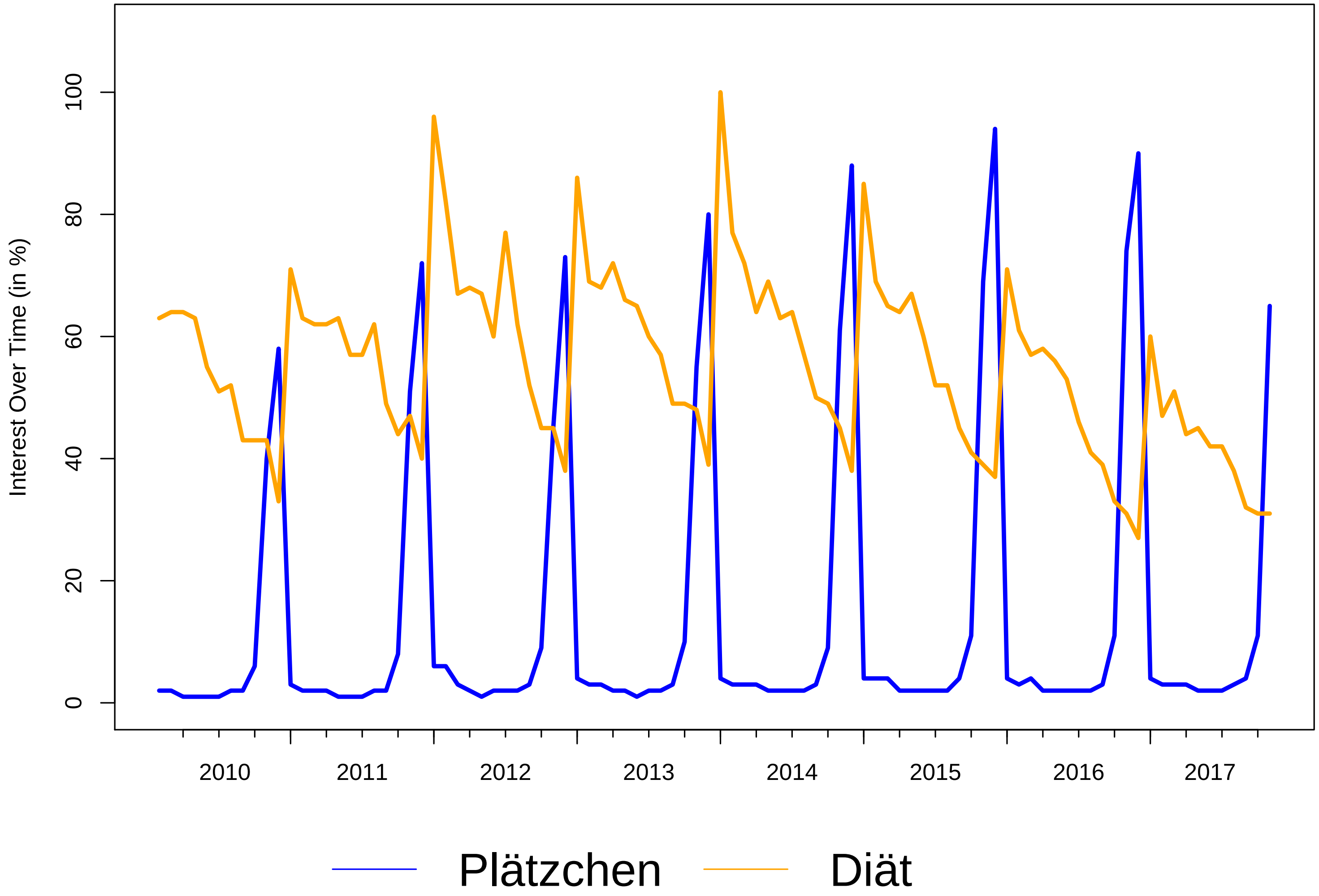





Source: Google Trends

# Internet searches reflect timing of individuals' actions



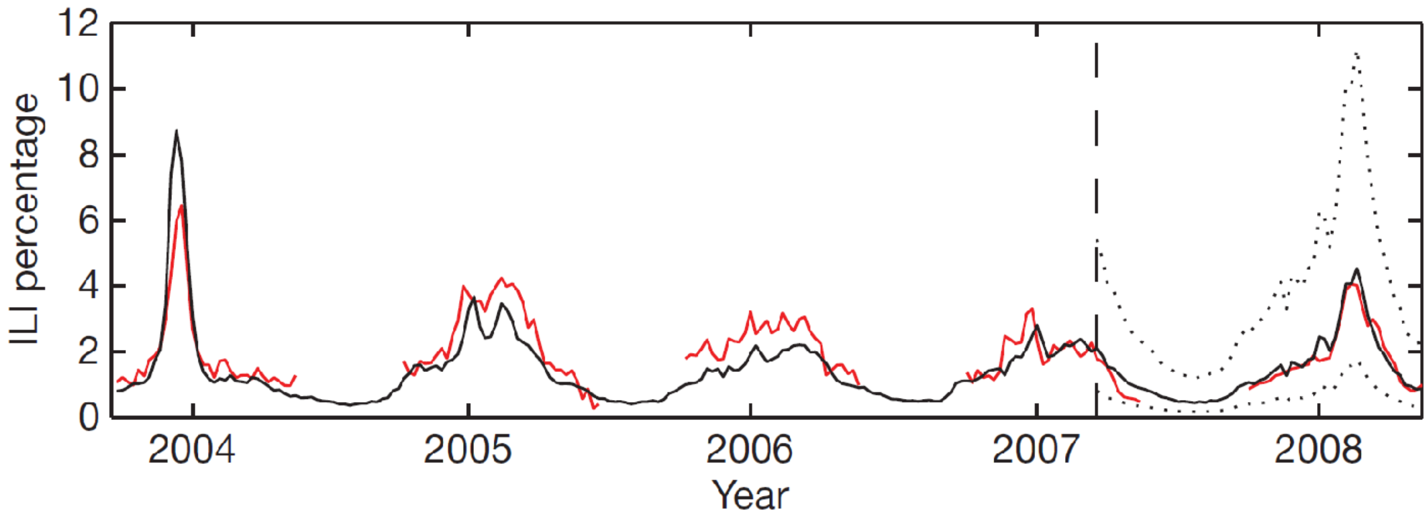



Source: Google Trends

# The first study that used google trends data predicted influenza epidemics

Ginsberg, Mohebbi, Patel, Brammer, Smolinski and Brilliant (2009, Nature)

## Influenza like illnesses (ILI)



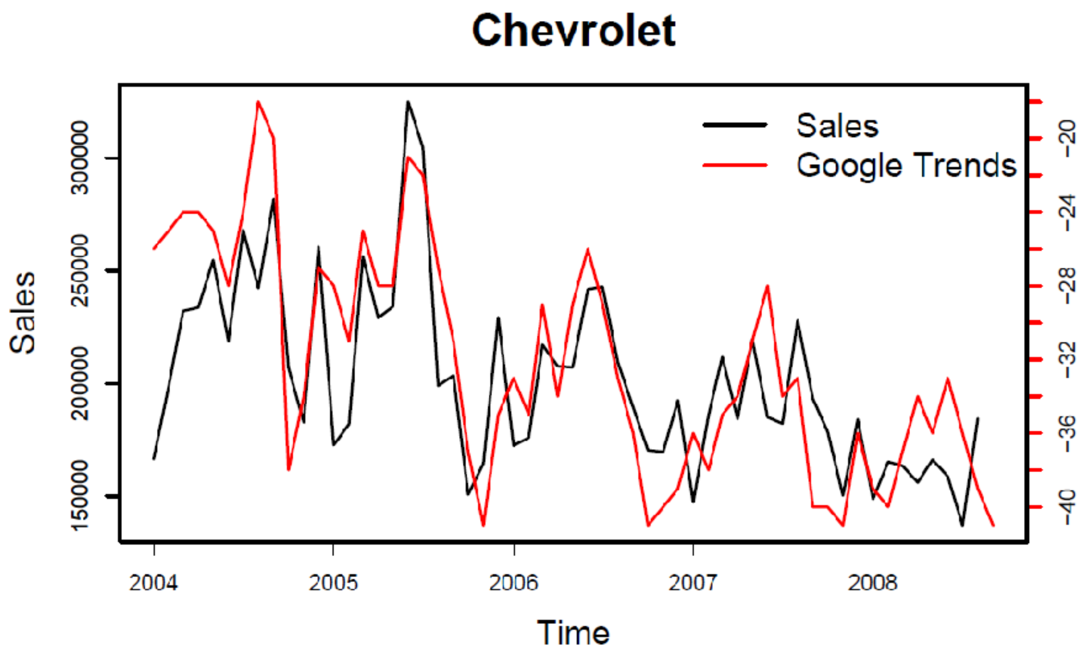

Reported (red), prediction(black)

*Source: Ginsberg et al. (2009), Figure 2*

# Google search volume data seem to carry information about what people are interested in



- Prediction of unemployment rates (Choi and Varian, 2009a)
- Prediction of retail sales (Choi and Varian, 2009b)




*Source: Choi & Varian (2009b), Figure 2.3*

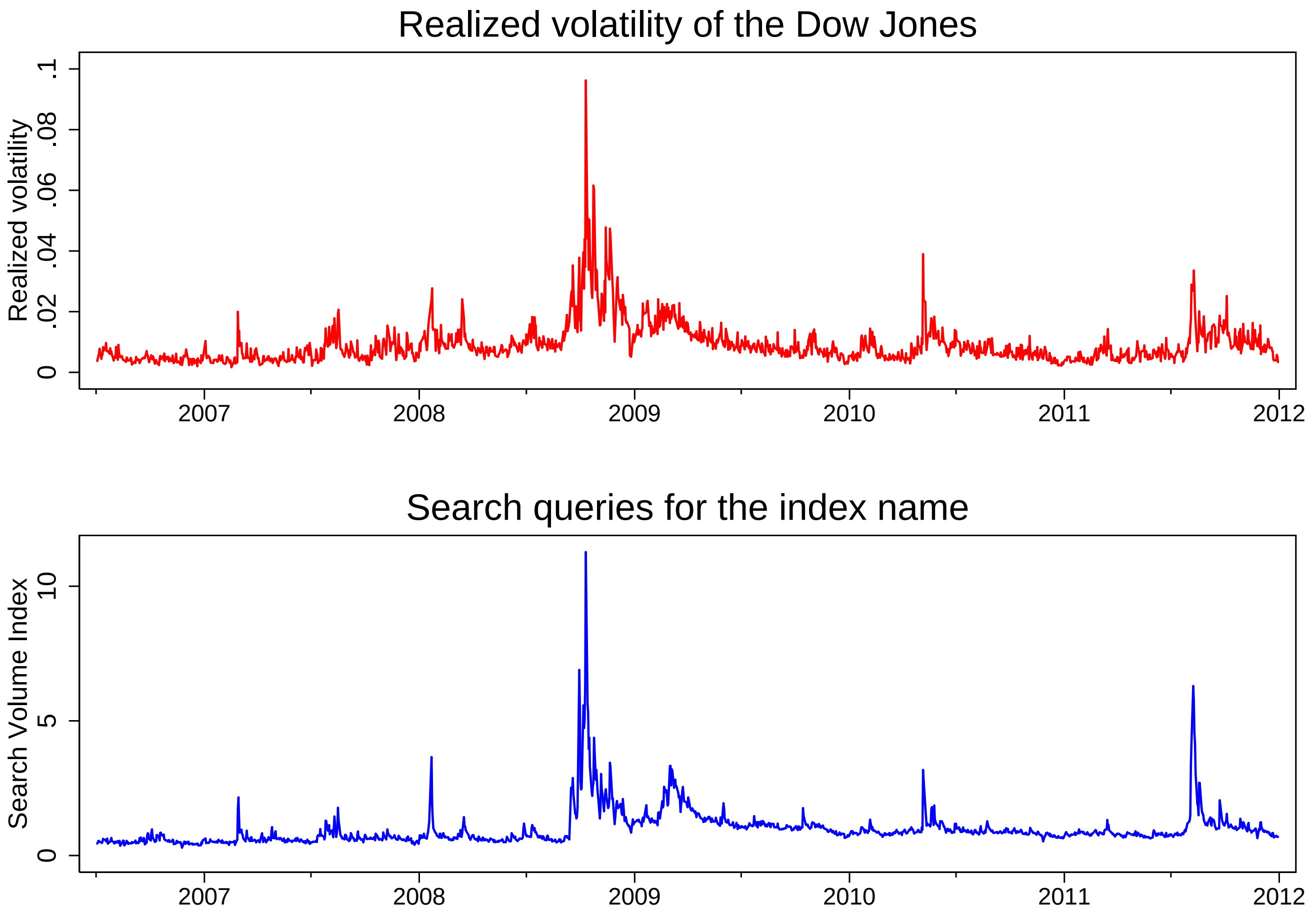
# There are numerous applications of Google search volume data in financial research



- Measure of retail investors' attention: individual stocks (Da, Engelberg, Gao, 2011, Journal of Finance)
- Weekly stock market volatility (Vlastakis and Markellos, 2012, Journal of Banking & Finance)
- The sum of all FEARS: Investor sentiment and asset prices (Da, Engelberg, Gao, 2014, The Review of Financial Studies)
- Investor Pessimism and the German Stock Market: Exploring Google Search Queries (Dimpfl and Kleiman, forthcoming, German Economic Review)
- Googling Gold and Mining Bad News (Dimpfl and Baur, 2016, Resources Policy)
- Can we predict the financial markets based on Google's search queries? (Perlin, Caldeira, Santos and Pontuschka, 2017, Journal of Forecasting)

# Large stock market movements capture investors' attention






Contemporaneous correlation (RV-SQ): $0.82$

# Why should searches be helpful to predict volatility?



Search queries proxy **retail** investors' attention/interest

Agent-based models of stock market volatility
Lux & Marchesi (1999, Nature):

- two agents: fundamentalists and noise traders
- fundamental price shock $\Rightarrow$ noise trading $\Rightarrow$ volatility

Recent evidence by Foucault et al. (2011, Journal of Finance): Noise traders contribute to volatility (approx. 23%)

# Dataset of realized volatilities and search queries

# What is “Realized volatility”



- simply put:
  RV is the daily variation of the price of a product, e.g. of a stock
- formally:
  RV is the (daily) standard deviation of the log returns of a stock
- Realized variance:

$$RVar_t = \sum_{i=0}^{n_t} r_{t,i}^2$$

with $n_t$ the number of intervals per day

- typical model for returns:

$$r_t = \sqrt{h_t}\,\eta_t$$

$$\text{with } \eta_t \ i.i.d.\ \mathcal{N}(0,1)$$

- On every day, prices are observed at every point in time $\tau_i, \tau_i \in \{\tau_0, ..., \tau_{n_t}\}$ during the day
  then $p_{t,i}$ ($i = 1, ..., n_t$) is the $i$-th observation on day $t$
- return between two intradaily points of time $i$ and $i-1$:

$$r_{t,i} = \sqrt{h_{t,i}}\,\eta_{t,i}$$

$$\text{with } \eta_{t,i} \ \ \mathcal{N}(0, \tfrac{1}{n_t})$$

$$r_{t,i} = p_{t,i} - p_{t,i-1}$$

- Hence, we get daily values as

$$r_t = \sum_{i=0}^{n_t} r_{t,i}$$

$$h_t = \frac{1}{n_t} \sum_{i=1}^{n_t} h_{t,i}$$

- Aim: show that $\mathbb{E}\left[r_t^2 | \mathcal{F}_{t,0}\right] = h_t$ holds

# Squared returns as a variance estimator

$$
\begin{aligned}
r_t^2 &= \left(\sum_{i=0}^{n_t} r_{t,i}\right)^2 \\
&= \sum_{i=0}^{n_t} r_{t,i}^2 + 2\sum_{i=0}^{n_t-1}\sum_{j=t+1}^{n_t} r_{t,i}r_{t,j} \\
\mathbb{E}\left[r_t^2|\mathcal{F}_0\right] &= \underbrace{\mathbb{E}\left[\sum_{i=0}^{n_t} r_{t,i}^2|\mathcal{F}_0\right]}_{RVar} \\
&+ 2\mathbb{E}\left[\sum_{i=0}^{n_t-1}\sum_{j=t+1}^{n_t} r_{t,i}r_{t,j}|\mathcal{F}_0\right]
\end{aligned}
$$

# Squared returns as a variance estimator



- Under the assumption that returns are uncorrelated, we get

$$\mathbb{E}\left[r_t^2|\mathcal{F}_{t,0}\right] = \mathbb{E}\left[RVar|\mathcal{F}_{t,0}\right] = h_t$$

- and finally

$$RV_t = \sqrt{RVar_t}$$

# The dataset

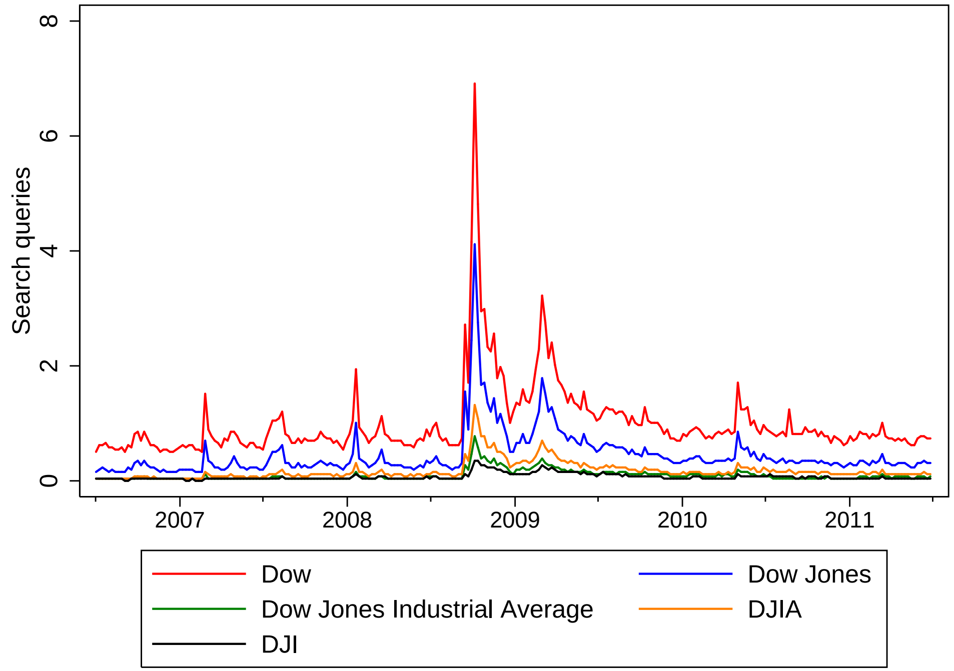


- Stock market index: Dow Jones
- Sample: daily data (trading days)
  5 1/2 years, July 2006 - December 2011
- Realized volatility (estimated on 10 min intervals)
- Search queries for index name in the US: “dow”
- Timing:
  - searches measured from 12 pm to 12 pm Pacific Standard Time
  - corresponds to 3 am Eastern Standard time, i.e. 6.5 hours before opening of the NYSE

# “Dow” is the appropriate search term

| Search terms correlated with “dow” | | | |
|---|---|---|---|
| Rank | Correlation | Scale | Term |
| 1. | 0.9981 | 46.67 | dow jones |
| 2. | 0.9707 | 18.40 | djia |
| 3. | 0.9672 | 6.13 | dow stock |
| 4. | 0.9612 | 0.33 | dow close |
| 5. | 0.9612 | 0.31 | current dow |
| 6. | 0.9591 | 9.91 | dow jones industrial |
| 7. | 0.9565 | 0.16 | current dow jones |
| 8. | 0.9545 | 0.39 | google dow |
| 9. | 0.9481 | 0.04 | stock market now |
| 10. | 0.9476 | 9.61 | industrial average |

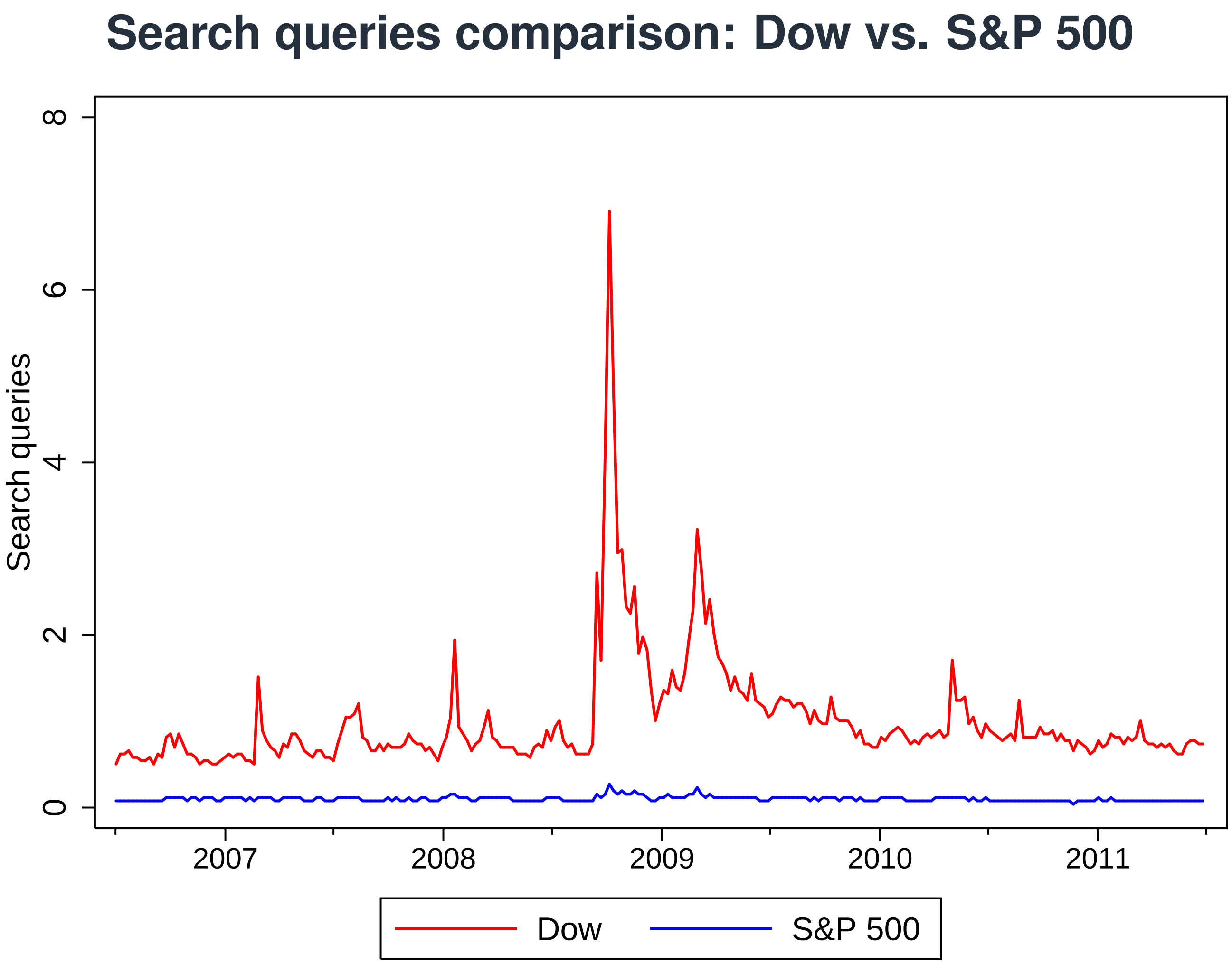
**Search queries comparison: Dow**




Issues with low search volume:

- Lower accuracy
- Missing values at daily frequency (threshold)

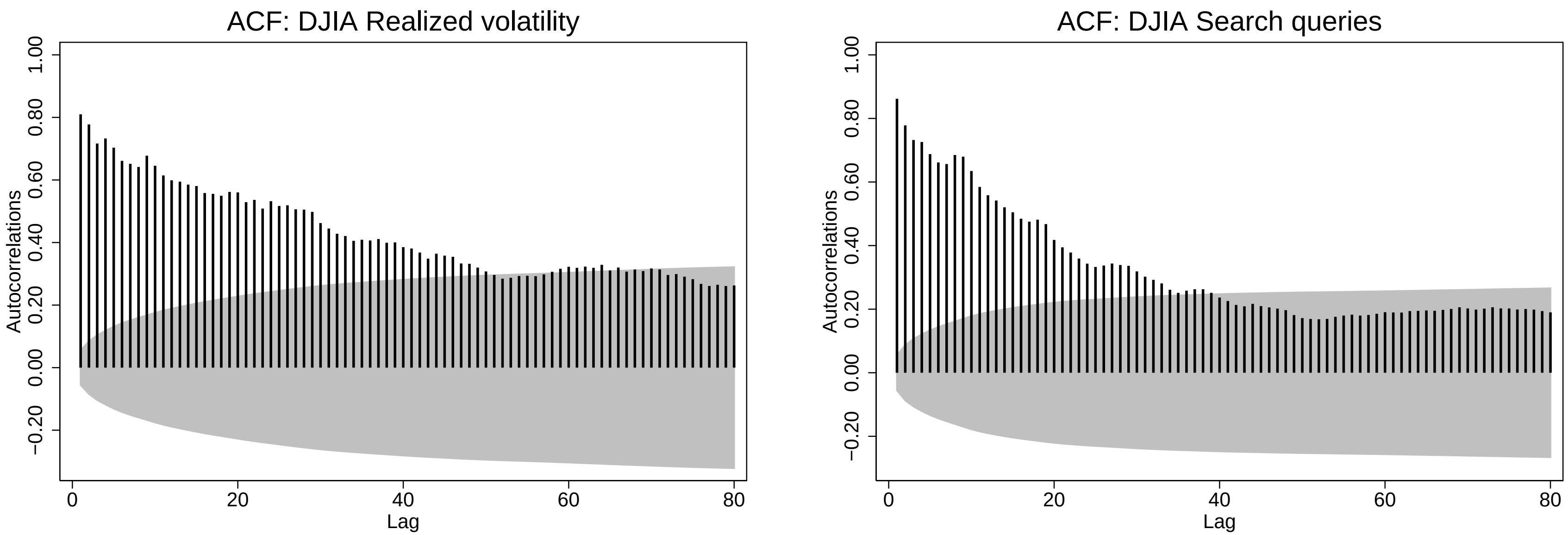
# Why we use the Dow and not the S&P 500




Weekly data, correlation ("Dow" - "S&P 500") $0.76$

# Autocorrelations of realized volatility and search queries





# Analysis of the joint dynamics in a vector autoregressive model

# Dynamics of realized volatility and search queries



**VAR estimation:**

| | log-RV$_t$ | log-SQ$_t$ |
|---|---|---|
| log-RV$_{t-1}$ | **0.42** | **0.03** |
| | (0.000) | (0.057) |
| log-RV$_{t-2}$ | **0.17** | -0.01 |
| | (0.000) | (0.582) |
| log-RV$_{t-3}$ | **0.09** | -0.02 |
| | (0.005) | (0.384) |
| log-RV$_{t-4}$ | **0.17** | 0.02 |
| | (0.000) | (0.352) |
| log-SQ$_{t-1}$ | **0.23** | **0.80** |
| | (0.000) | (0.000) |
| log-SQ$_{t-2}$ | -0.06 | -0.04 |
| | (0.347) | (0.275) |
| log-SQ$_{t-3}$ | -0.07 | **0.09** |
| | (0.299) | (0.021) |
| log-SQ$_{t-4}$ | 0.00 | **0.09** |
| | (0.993) | (0.002) |
| Constant | **-0.72** | 0.10 |
| | (0.000) | (0.139) |

**Granger causality test:**

| | log-RV | log-SQ |
|---|---|---|
| log-RV | | 5.97 |
| | | (0.201) |
| log-SQ | **30.47** | |
| | (0.000) | |

Granger-causality:

- SQ predicts RV

# Dynamics of search queries and trading volume



**VAR estimation:**

| | log-SQ$_t$ | log-VO$_t$ |
|---|---|---|
| log-SQ$_{t-1}$ | **0.80** | **0.20** |
| | (0.000) | (0.000) |
| log-SQ$_{t-2}$ | -0.04 | -0.08 |
| | (0.320) | (0.144) |
| log-SQ$_{t-3}$ | **0.10** | -0.01 |
| | (0.004) | (0.793) |
| log-SQ$_{t-4}$ | **0.09** | -0.08 |
| | (0.002) | (0.066) |
| log-VO$_{t-1}$ | 0.03 | **0.39** |
| | (0.185) | (0.000) |
| log-VO$_{t-2}$ | -0.02 | **0.20** |
| | (0.400) | (0.000) |
| log-VO$_{t-3}$ | **-0.05** | **0.13** |
| | (0.013) | (0.000) |
| log-VO$_{t-4}$ | 0.01 | **0.08** |
| | (0.587) | (0.005) |
| Constant | 0.17 | **1.12** |
| | (0.060) | (0.000) |

**Granger causality test:**

| | log-SQ | log-VO |
|---|---|---|
| log-SQ | | **25.24** |
| | | (0.000) |
| log-VO | **10.76** | |
| | (0.029) | |

Granger-causality:

- SQ predicts Volume
- Volume predicts SQ

# Forecast evaluation

## AR and VAR model of realized volatility (RV) and search queries (SQ):

$$log\text{-}RV_t = c_1 + \sum_{j=1}^{p} \beta_{1,j} log\text{-}RV_{t-j} + \gamma_{1,1} log\text{-}SQ_{t-1} + \varepsilon_{1,t}$$

$$log\text{-}SQ_t = c_2 + \beta_{2,1} log\text{-}RV_{t-1} + \sum_{j=1}^{q} \gamma_{2,j} log\text{-}SQ_{t-j} + \varepsilon_{2,t}.$$

## HAR and V-HAR model of realized volatility (RV) and search queries (SQ):

$$log\text{-}RV_t = c_1 + \beta_d log\text{-}RV_{t-1} + \beta_w log\text{-}RV^{w}_{t-1} + \beta_m log\text{-}RV^{m}_{t-1} + \gamma_{1,1} log\text{-}SQ_{t-1} + \varepsilon_{1,t}$$

$$log\text{-}SQ_t = c_2 + \beta_{2,1} log\text{-}RV_{t-1} + \sum_{j=1}^{q} \gamma_{2,j} log\text{-}SQ_{t-j} + \varepsilon_{2,t}.$$

- One-step ahead predictions: inclusion of SQ as an explanatory variable sufficient.
- Multi-step predictions: It is necessary to model and forecast SQ as well.

**Mean squared error and quasi likelihood loss function:**

$$\begin{aligned} \text{MSE} &= (RV_{t+1} - \widehat{RV}_{t+1|t})^2, \\ \text{QL} &= \frac{RV_{t+1}}{\widehat{RV}_{t+1|t}} - \log \frac{RV_{t+1}}{\widehat{RV}_{t+1|t}} - 1, \end{aligned}$$

(robust to possible noise in the volatility measure)

**Minzer-Zarnowitz Regression $\mathrm{R}^2$:**

$$RV_{t+1} = b_0 + b_1 \widehat{RV}_{t+1|t} + e_{t+1}.$$

| Model: | MSE | QL | $R^2$ |
|---|---|---|---|
| AR(1) | 0.172 | 22.82 | 65.96 |
| AR(1) + SQ | 0.160** | 22.10 | 67.66 |
| AR(4) | 0.147 | 21.18 | 70.65 |
| AR(4) + SQ | **0.145*** | **21.13** | **70.77** |
| HAR | 0.148 | 21.15 | 70.34 |
| HAR + SQ | 0.146* | **21.13** | 70.66 |

- HAR model: best among the univariate models
- Including SQ improves forecast for each univariate model
- HAR + SQ: best performing model

# Out-of-sample forecast evaluation

- Initial estimation window: 500 trading days (July 2006 to June 2008)
- Out-of-sample period: high volatility phase of financial crisis

| | | 1 day | | | 1 week | | | 2 weeks | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model: | | MSE | QL | $R^2$ | MSE | QL | $R^2$ | MSE | QL | $R^2$ |
| AR(1) | RV | 0.240 | 5.413 | 64.14 | 6.731 | 6.140 | 61.50 | 34.306 | 9.130 | 50.08 |
| VAR(1) | RV, SQ | 0.224** | 4.807 | 64.71 | 4.896** | 4.698 | 63.72 | 24.063** | 6.498 | 55.04 |
| AR(4) | RV | 0.199 | 4.352 | 69.36 | 3.977 | 3.633 | 70.94 | 18.220 | 4.734 | 65.28 |
| VAR(4) | RV, SQ | **0.191**** | 4.147 | **69.89** | 3.776** | 3.266 | 71.03 | 16.453** | 4.098 | 65.95 |
| HAR | RV | 0.198 | 4.324 | 69.07 | 3.655 | 3.450 | 72.07 | 15.699 | 4.198 | 67.74 |
| VHAR | RV, SQ | 0.194* | **4.142** | 69.74 | **3.568*** | **3.182** | **73.62** | **15.241*** | **3.768** | **70.82** |

**Intuition:**

- Shock of searches on volatility quite persistent.
- Good fit of time-series model for searches allows to iterate the system forward.

- Delayed publication of search query data (1 day)
- Focus of this paper: dynamics of volatility and attention
- Technically possible to publish search volume even faster (Google Hot Trends)

| Model: | | MSE | QL | $R^2$ |
|---|---|---|---|---|
| AR(1) | RV | 0.240 | 5.413 | 64.14 |
| VAR(1) | RV, SQ | 0.214** | 4.745 | 66.23 |
| AR(4) | RV | 0.199 | 4.352 | 69.36 |
| VAR(4) | RV, SQ | 0.196* | 4.253 | 69.32 |
| HAR | RV | 0.198 | 4.324 | 69.07 |
| VHAR | RV, SQ | **0.194** | **4.223** | **69.43** |

# Forecast evaluation across different times of volatility: in-sample prediction

| | | Realized Volatility – Bottom Quantiles | | | | |
|---|---|---|---|---|---|---|
| | | 1% | 5% | 10% | 25% | 50% |
| MSE: | HAR | 0.040 | 0.024 | 0.021 | 0.020 | 0.027 |
| MSE: | HAR + SQ | 0.038 | 0.024 | 0.021 | 0.020 | 0.027 |
| Difference in MSE | | 0.002 | 0.000 | 0.000 | 0.000 | 0.001 |
| QL | HAR | 15.23 | 7.84 | 5.70 | 4.07 | 3.54 |
| QL | HAR + SQ | 15.01 | 7.90 | 5.74 | 4.04 | 3.46 |
| Difference in QL | | 0.22 | -0.05 | -0.04 | 0.03 | 0.07 |

| | | Realized Volatility – Top Quantiles | | | | |
|---|---|---|---|---|---|---|
| | | 50% | 25% | 10% | 5% | 1% |
| MSE: | HAR | 0.267 | 0.487 | 1.047 | 1.870 | 6.875 |
| MSE: | HAR + SQ | 0.263 | 0.479 | 1.018 | 1.804 | 6.560 |
| Difference in MSE | | 0.005 | 0.008 | 0.029 | 0.067 | 0.315 |
| QL | HAR | 5.51 | 7.53 | 11.06 | 13.73 | 36.36 |
| QL | HAR + SQ | 5.52 | 7.59 | 10.93 | 13.15 | 34.21 |
| Difference in QL | | -0.01 | -0.06 | 0.13 | 0.58 | 2.15 |

# Forecast evaluation across different times of volatility: Out-of-sample forecast



| | | Realized Volatility – Bottom Quantiles | | | | |
|---|---|---|---|---|---|---|
| | | 1% | 5% | 10% | 25% | 50% |
| MSE: | HAR | 0.016 | 0.026 | 0.028 | 0.029 | 0.034 |
| MSE: | VHAR (RV,SQ) | 0.020 | 0.029 | 0.031 | 0.031 | 0.036 |
| Difference in MSE | | -0.003 | -0.003 | -0.003 | -0.002 | -0.002 |
| QL | HAR | 6.63 | 6.29 | 5.51 | 3.91 | 3.50 |
| QL | VHAR (RV,SQ) | 7.74 | 7.01 | 6.01 | 4.18 | 3.56 |
| Difference in QL | | -1.11 | -0.72 | -0.50 | -0.27 | -0.06 |

| | | Realized Volatility – Top Quantiles | | | | |
|---|---|---|---|---|---|---|
| | | 50% | 25% | 10% | 5% | 1% |
| MSE: | HAR | 0.362 | 0.664 | 1.445 | 2.549 | 7.554 |
| MSE: | VHAR (RV,SQ) | 0.351 | 0.644 | 1.338 | 2.340 | 6.691 |
| Difference in MSE | | 0.010 | 0.020 | 0.107 | 0.209 | 0.863 |
| QL | HAR | 5.15 | 6.93 | 10.98 | 16.85 | 21.41 |
| QL | VHAR (RV,SQ) | 4.72 | 6.30 | 9.09 | 13.59 | 15.44 |
| Difference in QL | | 0.42 | 0.63 | 1.88 | 3.26 | 5.96 |

# Economic Value of Volatility Timing

# or:

# can you earn money with this stuff?

# Economic Value of Volatility Timing

Quadratic utility of an investor:

$$U(W_{t+1}) = W_t \, r_{p,t+1} - \frac{aW_t^2}{2} \, r_{p,t+1}^2$$

- $r_{p,t+1}$: return of the investor's portfolio
- $a$ absolute risk aversion
- Portfolio consists of a risky asset (market portfolio) and a risk-free asset

Investor implements a variance targeting strategy ($\sigma_p^2 = 12\%$):

- In $t$: investor derives weights
- In $t+1$: portfolio gains and losses are realized
- Variance prediction based on either HAR model of realized volatility or HAR model including Google search queries
- Compare managed investment strategies to static buy and hold

Assessment of economic gains:

- $r_{A,t}$ and $r_{B,t}$: returns of two alternative portfolios
- $\Delta$: maximum fee investor is willing to pay to switch from portfolio A to portfolio B

Compare utility of the two investments to find $\Delta$:

$$\sum_{t=0}^{T-1} \left[ (r_{A,t+1} - \Delta) - \frac{\gamma}{2(1+\gamma)} (r_{A,t+1} - \Delta)^2 \right]$$
$$= \sum_{t=0}^{T-1} \left[ r_{B,t+1} - \frac{\gamma}{2(1+\gamma)} r_{B,t+1}^2 \right]$$

# Economic Value of Volatility Timing



Additional gain from including Google search queries:

- Different levels of relative risk aversion $\gamma$
- Performance fee paid for search query based portfolio is robust with respect to risk aversion

| | relative risk aversion $\gamma =$ | | |
|---|---|---|---|
| | 1 | 5 | 10 |
| $\Delta$(50-HAR) | 112.05 | 110.25 | 109.84 |
| $\Delta$(50-HAR+SQ) | 121.14 | 119.60 | 119.24 |
| $\Delta$(HAR-HAR+SQ) | 8.99 | 9.24 | 9.30 |

# Conclusion

- VAR analysis: Results are consistent with an agent-based/noise trader theory of (additional) volatility:
  - Co-movement of volatility and retail investors' attention
  - Search queries predict volatility
  - Search queries predict volume
- Search queries are a valuable source of information for future volatility.
- Forecast improvements ...
  - in-sample
  - out-of-sample
  - longer forecast horizons
  - esp. in high volatility phases (e.g. the financial crisis of 2008)

# Thank you.

# References

Baur, D. G., Dimpfl, T., 2016. Googling gold and mining bad news. Resources Policy 50, 306–311

Choi, H., Varian, H., 2009a. Predicting initial claims for unemployment benefits. Working paper

Choi, H., Varian, H., 2009b. Predicting the present with Google trends. Working paper

Da, Z., Engelberg, J., Gao, P., 2011. In Search of Attention. The Journal of Finance 66, 1461–1499

Da, Z., Engelberg, J., Gao, P., 2015. The Sum of All FEARS Investor Sentiment and Asset Prices. Review of Financial Studies 28, 1–32

Dimpfl, T., Kleiman, V., 2017. Investor Pessimism and the German Stock Market: Exploring Google Search Queries. German Economic Review doi: 10.1111/geer.12137

Foucault, T., Sraer, D., Thesmar, D. J., 2011. Individual Investors and Volatility. The Journal of Finance 66, 1369–1406

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–1014

Lux, T., Marchesi, M., 1999. Scaling and criticality in a stochastic multi-agent model of a financial market. Nature 397, 498–500

Perlin, M. S., Caldeira, J. a. F., Santos, A. A. P., Pontuschka, M., 2017. Can We Predict the Financial Markets Based on Google's Search Queries? Journal of Forecasting 36, 454–467

Vlastakis, N., Markellos, R. N., 2012. Information demand and stock market volatility. Journal of Banking & Finance 36, 1808 – 1821