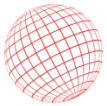
# Big data in macroeconomics

Lucrezia Reichlin
London Business School and now-casting economics ltd

COEURE workshop
Brussels 3-4 July 2015

Now-Casting.com
economics in real time

# *WHAT IS BIG DATA IN ECONOMICS?*

Frank Diebold claimed to have introduced the term in econometrics and statistics

*"I stumbled on the term Big Data innocently enough, via discussion of two papers that took a new approach to macro-econometric dynamic factor models (DFMs), Reichlin (2003) and Watson (2003), presented back-to-back in an invited session of the 2000 World Congress of the Econometric Society"*

# *3 IDEAS IN THE RESEARCH PROGRAM THAT WATSON AND I PRESENTED AT THE 2000 SEATTLE WORLD CONGRESS*

1. There is more data around than what is exploited in standard macro models

   Examples of data potentially interesting – but many more ….

   -- micro data: does heterogeneity matter?

   -- conjunctural indicators – used in long tradition of understanding/dating business cycles – typically available at higher frequency than national account, possibly more timely

2. Economic data are correlated (business cycle)

3. In developing models including large number of series need to understand what one is capturing when increasing the sample size in both n and t dimension: How should we think about convergence of estimators in this world? How should we think of the principle of parsimony in this world? [n,T asymptotic]

UNDERSTANDING THESE ISSUES IS STILL UNFINISHED BUSINESS BOTH FROM A THEORY AND EMPIRICAL POINT OF VIEW

## *THE PROBLEM OF THE CURSE OF DIMENSIONALITY*

- In large models there is a proliferation of parameters that is likely to lead to high estimation uncertainty
- As we increase complexity, the number of parameters to estimate increases and so does the variance (estimation uncertainty)
- Predictions based on traditional methods are poor or unfeasible if the number of predictors (n) is large relative to the sample size (T)

*Why?*

- The sample variance is inversely proportional to the degrees of freedom - sample size minus no. of parameters
- When no. of parameters becomes large, the d.f. go to zero or become negative – precision of the estimates deteriorate

*This is the curse of dimensionality!*

## *FACTOR MODELS WAS THE SOLUTION WE STUDIED AT THE TIME …*

*Insight of early work*

- curse of dimensionality problem can be solved if there are few common sources of variation in the data
- limit complexity due to proliferation of parameters by focusing on few sources of variations (common factors)
- Reasonable if data are characterized by strong collinearity: eg business cycles

# *Correlation in macro data*
# *an old insight from the 1970s*
# (from about 10 series to about 100)

Fraction of Variance Explained by 1- and 2-Factor Models

| Series | Sargent and Sims | | Giannone-Reichlin-Sala | |
|---|---|---|---|---|
| | 1 Factor | 2 Factors | 1 Factor | 2 Factors |
| Avg. Weekly Hours | 0.77 | 0.80 | 0.49 | 0.61 |
| Layoffs | 0.83 | 0.85 | 0.72 | 0.82 |
| Employment | 0.86 | 0.88 | 0.85 | 0.91 |
| Unemployment | 0.77 | 0.85 | 0.74 | 0.82 |
| Industrial Production | 0.94 | 0.94 | 0.88 | 0.93 |
| Retail Sales | 0.46 | 0.69 | 0.33 | 0.47 |
| New Orders Durables | 0.67 | 0.86 | 0.65 | 0.74 |
| Sensitive Material Prices | 0.19 | 0.74 | 0.53 | 0.60 |
| Wholesale Prices | 0.20 | 0.69 | 0.34 | 0.67 |
| M1 | 0.16 | 0.20 | 0.15 | 0.30 |
| Net Bus. Formation | 0.42 | 0.46 | NA | NA |

## *INFERENCE IN LARGE MODELS – n, T ASYMPTOTIC*

- The problem:

*how many variables can we handle for a given sample size? How does the estimate behaves as the no. of parameters increase?*

Questions that could not be handled with traditional approach to asymptotic which keeps the number of parameters fixed

In a series of papers we developed a new approach to asymptotic in the n and T dimension

We made a first step: under some conditions (co-movements) we can estimate large models consistently - Derive consistency results for n,T $\rightarrow \infty$ in large factor models

- *Estimators: principal components – early work*
- *Later: show n,T consistency for quasi max likelihood*

# *But why only considering factor models?*

Factor analysis is a particular way to limit parameter estimation uncertainty

Alternative idea: limit estimation uncertainty via shrinkage –

Penalized regression:

$$\min[\ \text{RSS(model)} + \lambda\ (\text{Model Complexity})]$$

Reduce variance and introduce bias

# *Bayesian is natural way to go …*

Penalized regression can be reinterpreted a Bayesian regression with normal prior

The Bayesian Solution: Mixed Estimation (Stein, 1956)

| Data (Complex/Rich) | + | Prior (Parsimonious/Naive) |
|---|---|---|

Stable and reliable estimation of complex and large model if dimension of the problem is finite [again comovements]

De Mol, Giannone and Reichlin, JoE 2008:

- (n,T) consistency at any rates – normal prior single regression
- Study empirically forecasting performance of Lasso, ridge and PC

# *INSIGHT*
# when data are correlated (macro) alternative methods have similar performance

Example: forecasting industrial production

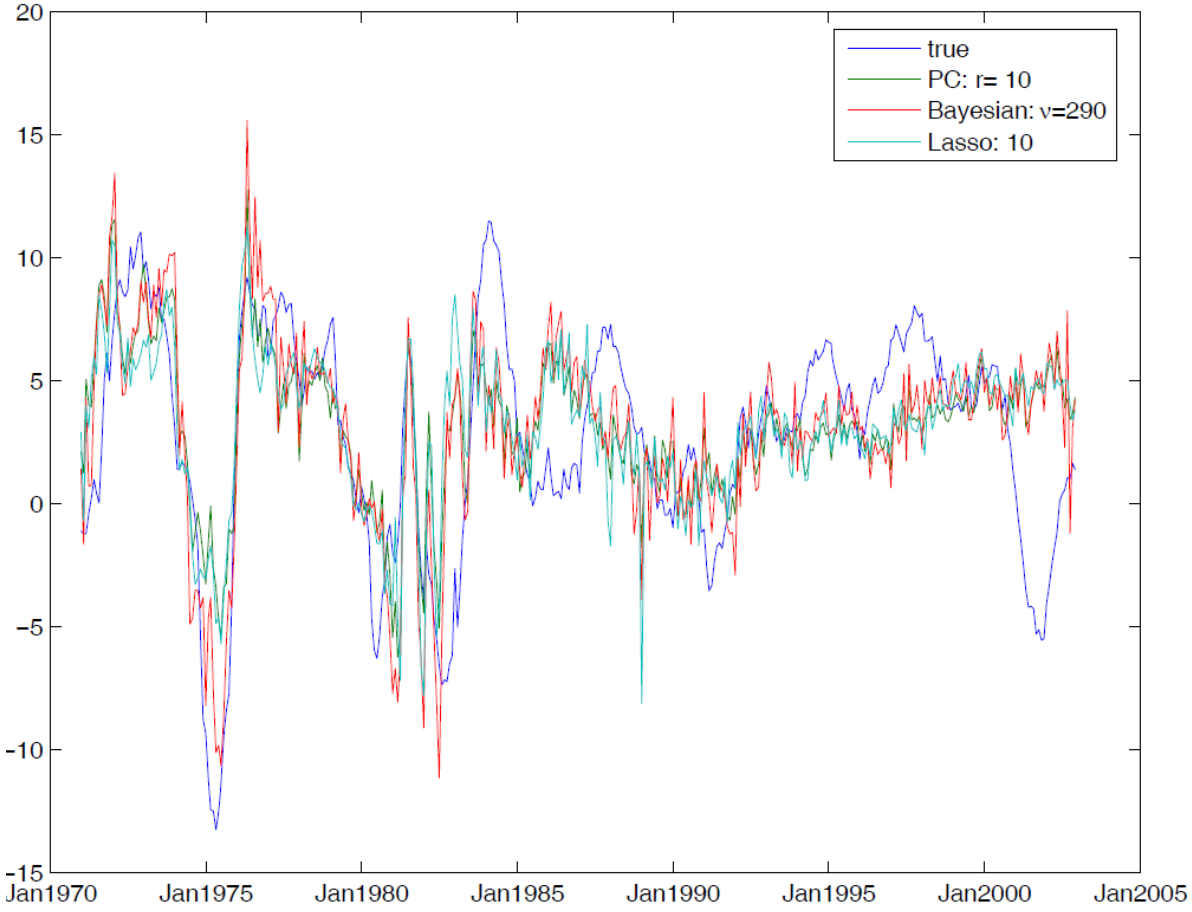
131 monthly series for the US economy

Consider:

- Ridge regression
- Lasso regression
- Principal component

Results: forecasts correlated, performance similar, Lasso variable election unstable

# Alternative forecasts (130 predictors)



# *Large Bayesian VARs*

Bayesian regression in a dynamic system of simultaneous equations had been applied in macro for small models since the 80s

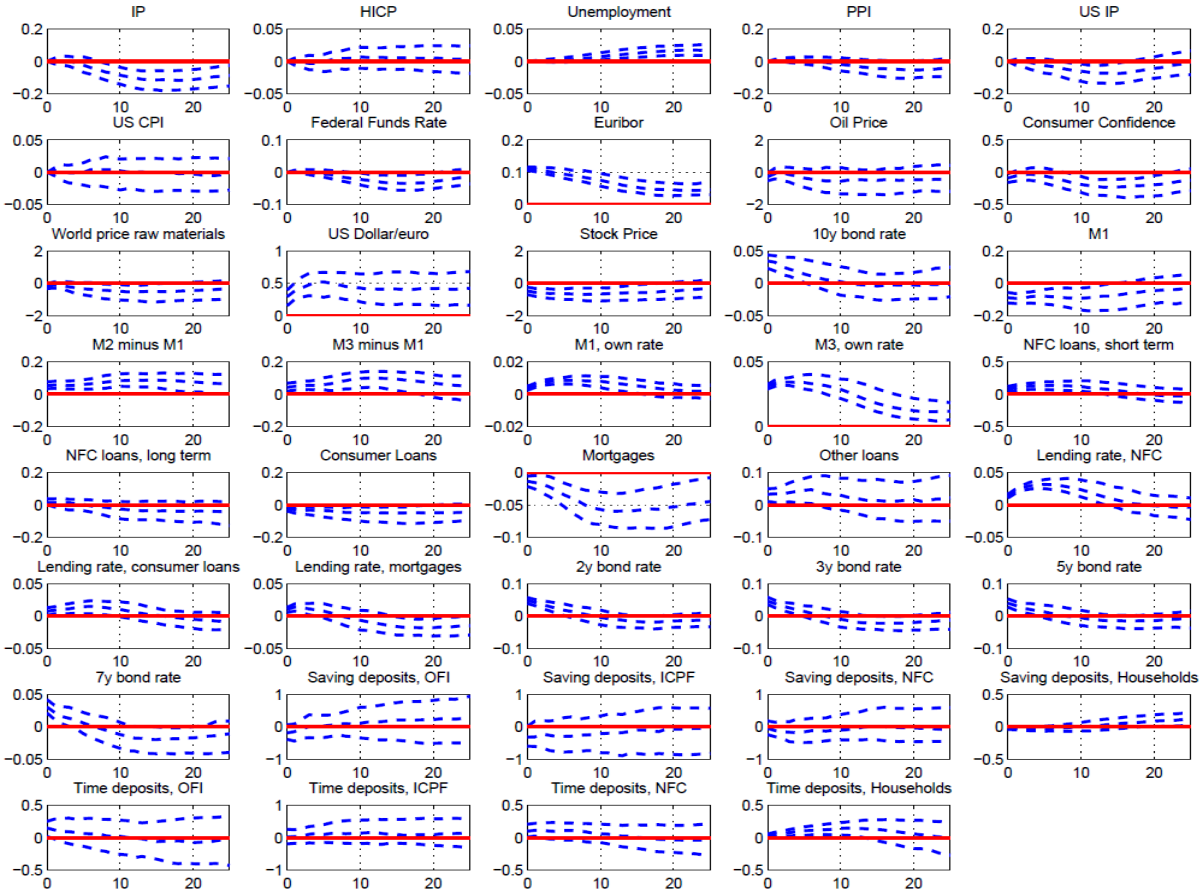
(B-VAR literature a la Doan, Litterman and Sims)

But results in De Mol et al. suggested that one could estimate VAR with many variables

By shrinking appropriately (in relation to the sample size) one can learn from the data and avoid over-fitting

Many successful applications: impulse response functions, counterfactuals, stress tests

# *Monetary VAR with 40 variables*
# *Much used for policy analysis at the ECB*



# *More empirical experience with big data in economics*

- **One of the most successful application of big data in economics has been *now-casting*: the real time monitoring of the "rich" data flow**
- Basic idea of now-casting:
  - follow the calendar of data publication
  - update now-cast almost in continuous time
  - corresponding to each release there will be a model based "surprise" that move the now-cast of all variables and the synthetic signal on the state of the economy

- THIS IS WHAT THE MARKET INFORMALLY DOES!

# Conjuctural information: This Week

| Jun 17 - Jun 23 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Date | 10:07am | Currency | Impact | | Detail | Actual | Forecast | Previous |
| Sun Jun 17 | | | | | | | | |
| Mon Jun 18 | 4:00pm | USD | | NAHB Housing Market Index | | 29 | 28 | 28 |
| | Day 1 | ALL | | G20 Meetings | | | | |
| Tue Jun 19 | 2:30pm | USD | | Building Permits | | 0.78M | 0.73M | 0.72M |
| | 2:30pm | USD | | Housing Starts | | 0.71M | 0.72M | 0.74M |
| | Day 2 | ALL | | G20 Meetings | | | | |
| Wed Jun 20 | 4:30pm | USD | | Crude Oil Inventories | | 2.9M | -1.0M | -0.2M |
| | 6:32pm | USD | | FOMC Statement | | | | |
| | 6:32pm | USD | | Federal Funds Rate | | <0.25% | <0.25% | <0.25% |
| | 8:00pm | USD | | FOMC Economic Projections | | | | |
| | 8:15pm | USD | | FOMC Press Conference | | | | |
| Thu Jun 21 | 2:30pm | USD | | Unemployment Claims | | 387K | 381K | 389K |
| | 3:00pm | USD | | Flash Manufacturing PMI | | 52.9 | 53.4 | 54.0 |
| | 4:00pm | USD | | Existing Home Sales | | 4.55M | 4.58M | 4.62M |
| | 4:00pm | USD | | Philly Fed Manufacturing Index | | -16.6 | 0.7 | -5.8 |
| | 4:00pm | USD | | CB Leading Index m/m | | 0.3% | 0.2% | -0.1% |
| | 4:00pm | USD | | HPI m/m | | 0.8% | 0.5% | 1.6% |
| | 4:30pm | USD | | Natural Gas Storage | | 62B | 64B | 67B |

# Conjuctural information: Next Week

| Jun 24 - Jun 30 | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Date | 10:08am | Currency | Impact | | Detail | Actual | Forecast | Previous |
| Sun Jun 24 | | | | | | | | |
| Mon Jun 25 | 4:00pm | USD | | New Home Sales | | | 347K | 343K |
| Tue Jun 26 | 3:00pm | USD | | S&P/CS Composite-20 HPI y/y | | | -2.4% | -2.6% |
| | 4:00pm | USD | | CB Consumer Confidence | | | 64.0 | 64.9 |
| | 4:00pm | USD | | Richmond Manufacturing Index | | | 5 | 4 |
| Wed Jun 27 | 2:30pm | USD | | Core Durable Goods Orders m/m | | | 1.0% | -0.9% |
| | 2:30pm | USD | | Durable Goods Orders m/m | | | 0.5% | 0.0% |
| | 4:00pm | USD | | Pending Home Sales m/m | | | 1.3% | -5.5% |
| | 4:30pm | USD | | Crude Oil Inventories | | | | 2.9M |
| Thu Jun 28 | 2:30pm | USD | | Unemployment Claims | | | 385K | 387K |
| | 2:30pm | USD | | Final GDP q/q | | | 1.9% | 1.9% |
| | 2:30pm | USD | | Final GDP Price Index q/q | | | 1.7% | 1.7% |
| | 4:30pm | USD | | Natural Gas Storage | | | | 62B |
| | 5:30pm | USD | | FOMC Member Pianalto Speaks | | | | |
| Fri Jun 29 | 2:30pm | USD | | Core PCE Price Index m/m | | | 0.2% | 0.1% |
| | 2:30pm | USD | | Personal Spending m/m | | | 0.1% | 0.3% |
| | 2:30pm | USD | | Personal Income m/m | | | 0.3% | 0.2% |
| | 3:45pm | USD | | Chicago PMI | | | 53.1 | 52.7 |
| | 3:55pm | USD | | Revised UoM Consumer Sentiment | | | 74.3 | 74.1 |

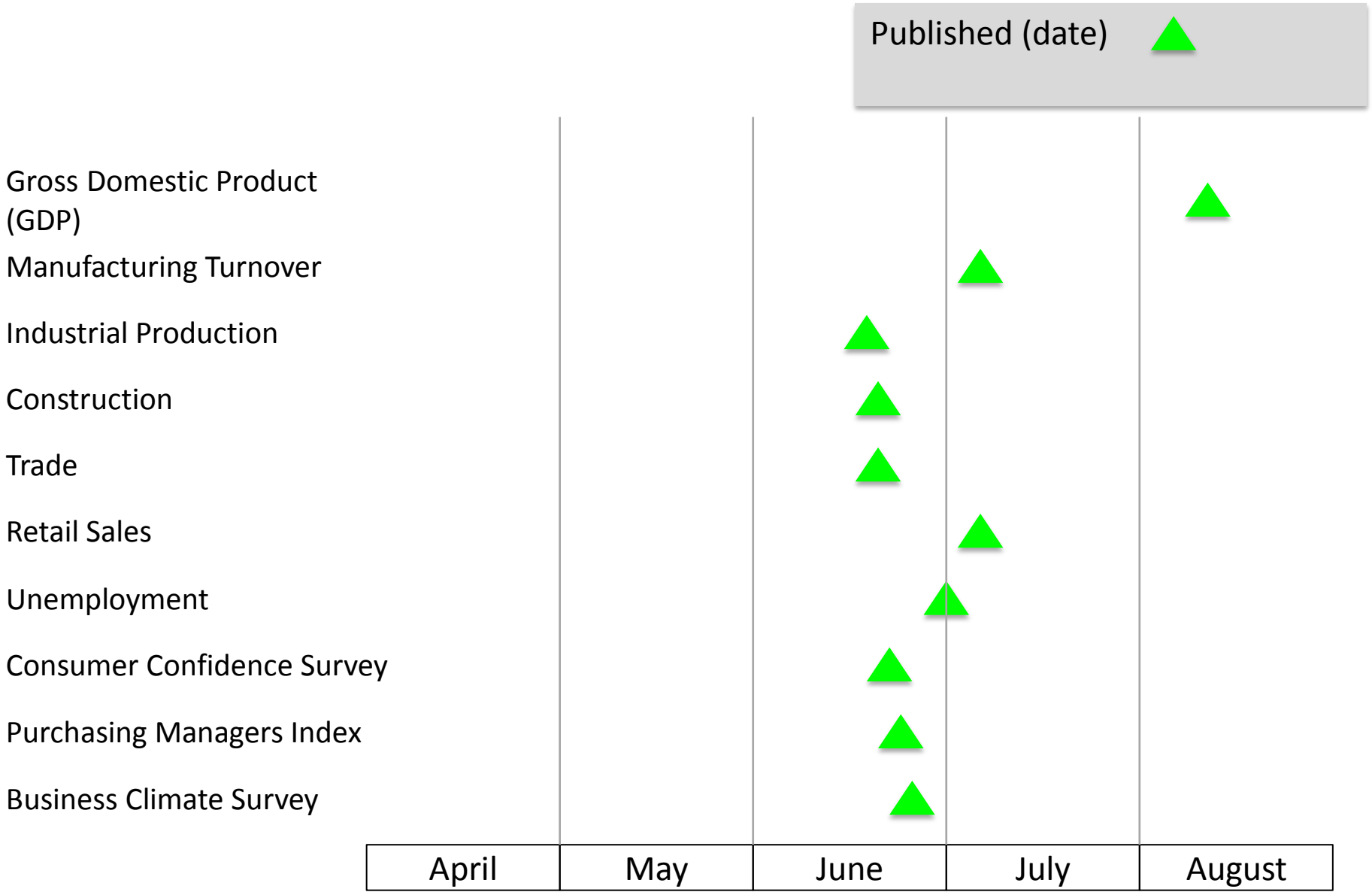
# *The structure of the problem is non standard*

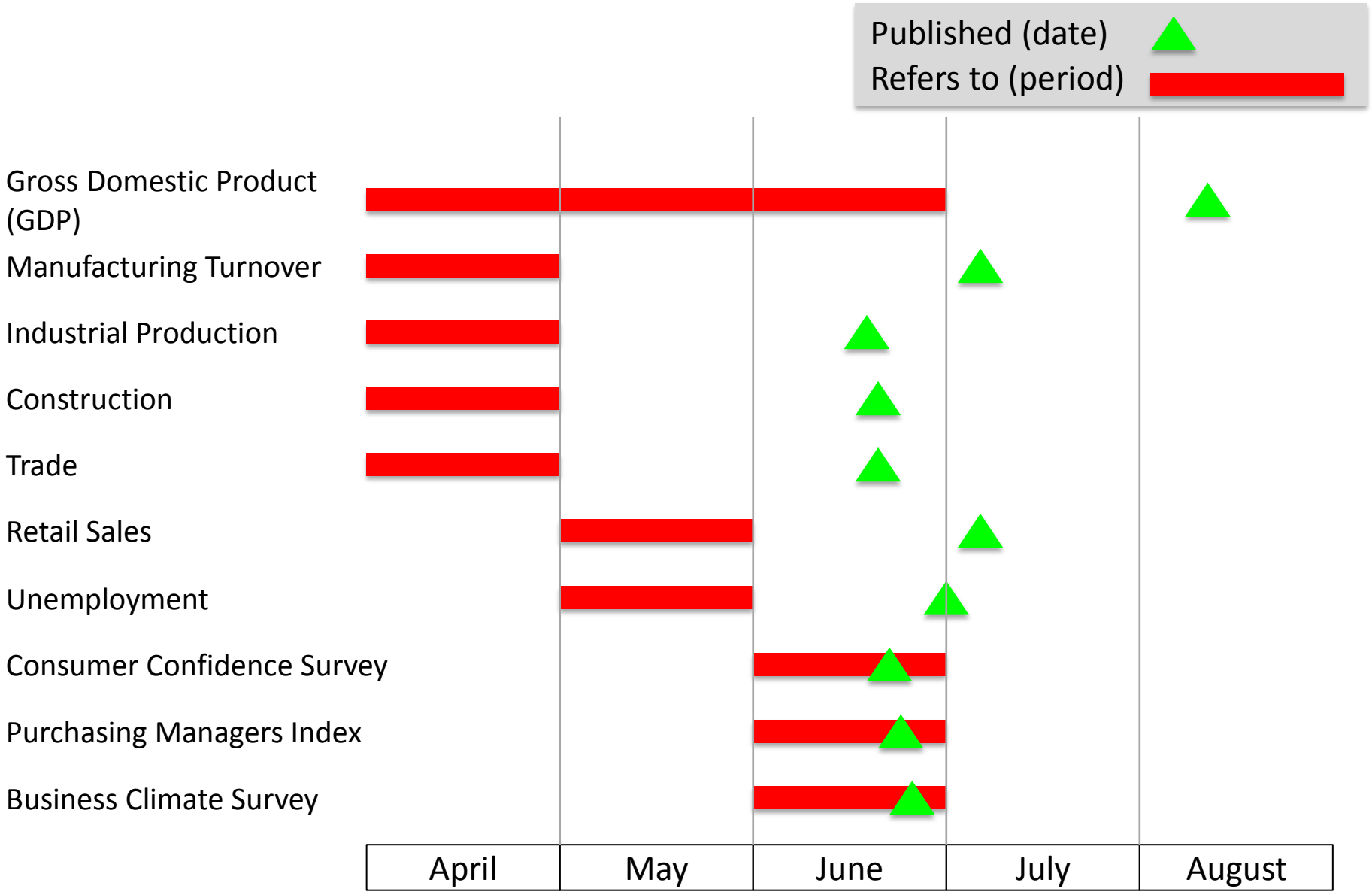
It is a big data problem but in addition:

1. Mixed frequency and time aggregation for stocks and flow variables
2. Non synchronous calendar
3. Missing observations

Problems largely solved by recent research: see our Holland Handbook chapter on now-casting

- Many available data – such as surveys - are valuable because of their timeliness
- But need to understand details of the structure of the information flow problem
- Cannot simply borrowing from other disciplines!

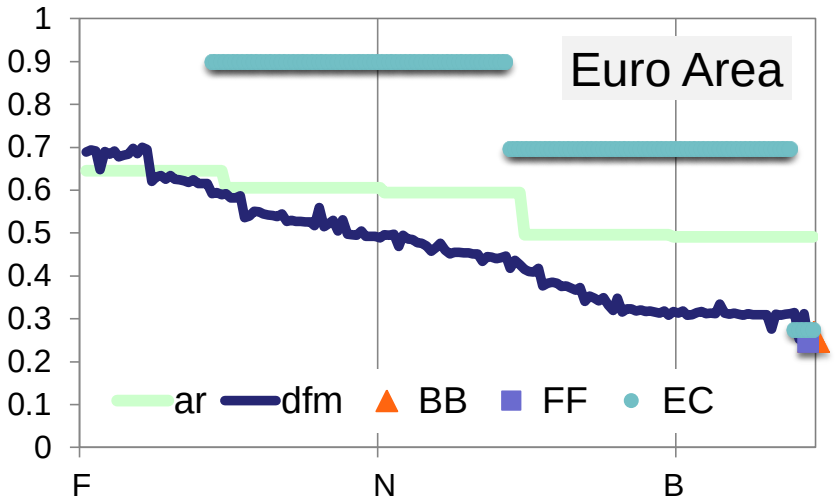

Published (date)
Gross Domestic Product (GDP)
Manufacturing Turnover
Industrial Production
Construction
Trade
Retail Sales
Unemployment
Consumer Confidence Survey
Purchasing Managers Index
Business Climate Survey
April
May
June
July
August

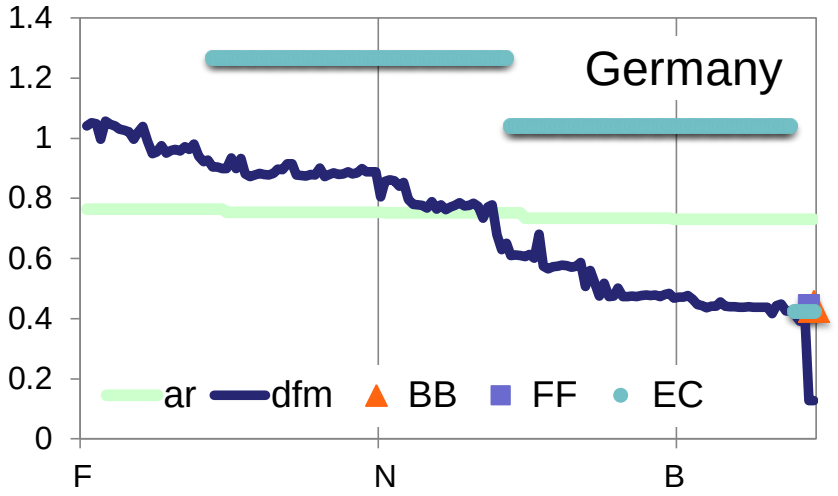

Published (date)
Refers to (period)
Gross Domestic Product (GDP)
Manufacturing Turnover
Industrial Production
Construction
Trade
Retail Sales
Unemployment
Consumer Confidence Survey
Purchasing Managers Index
Business Climate Survey
April
May
June
July
August

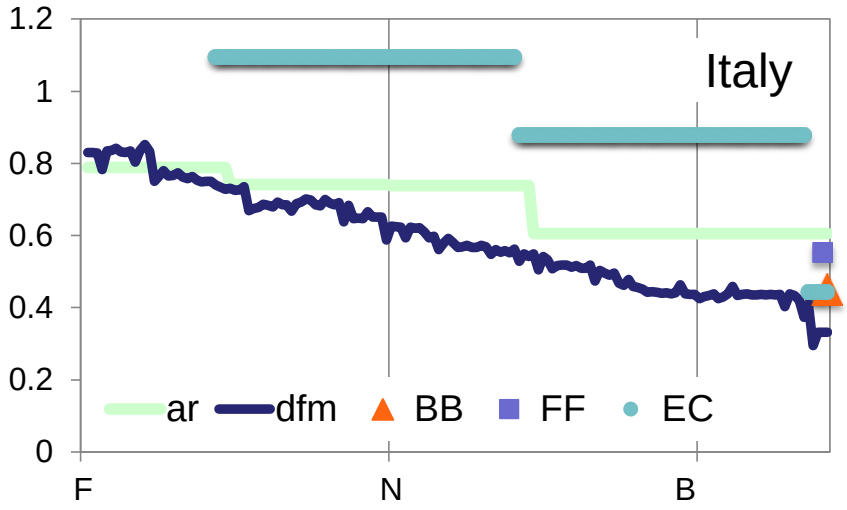
# *What have learned in years of experience?*

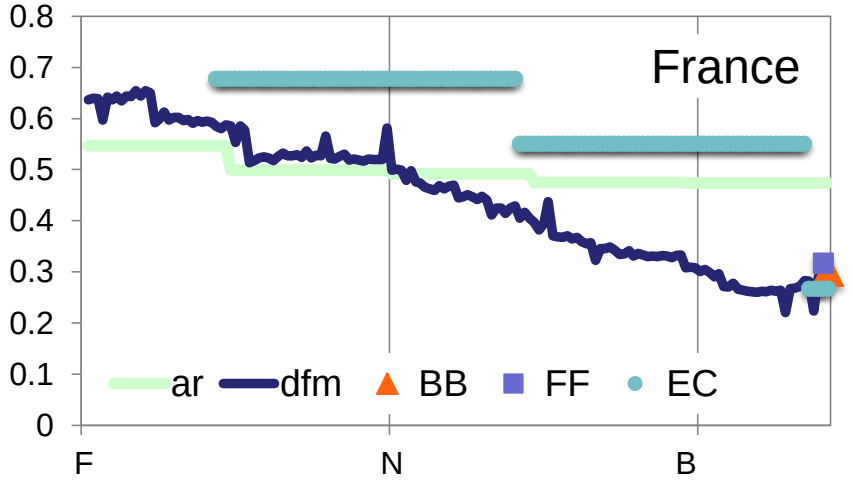
- Timeliness matters
- Many data are relevant to obtain early signals on economic activity, increasingly also used by statistical agencies
- In particular: surveys, weekly conjunctural data
- Robust models are relatively simple
- An automatic mechanical model does as well as judgment but is as timely as you want and does not get influenced by moods

# *Do timely data help? Evolution of the MSFE in relation to the data flow*

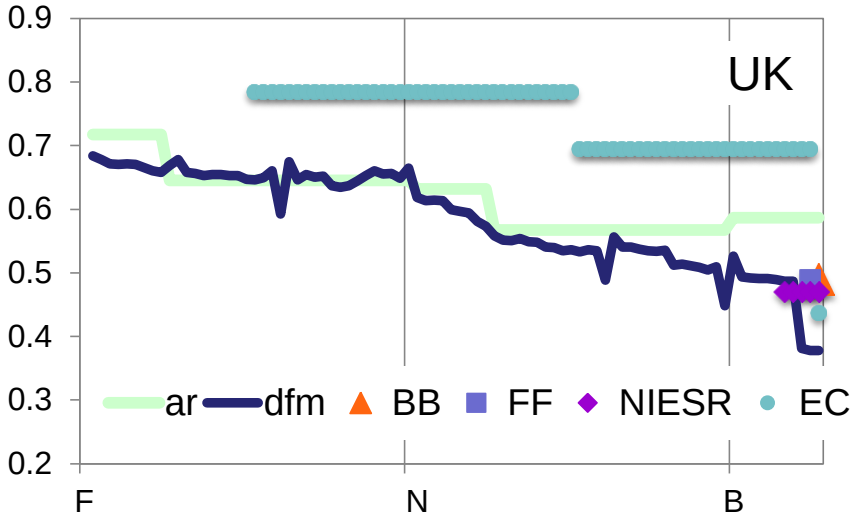


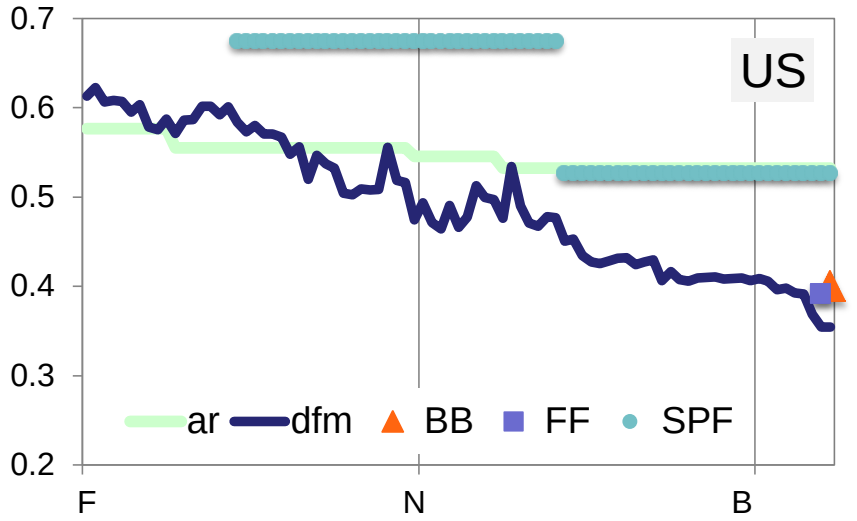



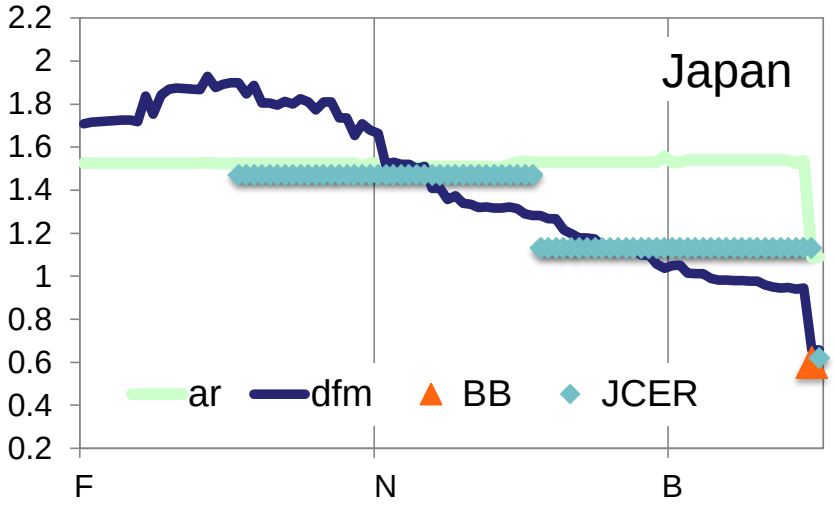



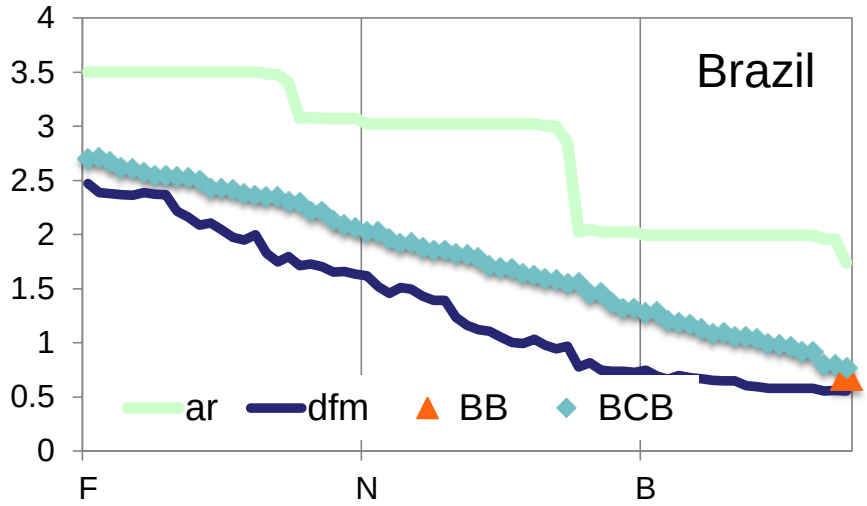



# *Do timely data help? Evolution of the MSFE in relation to the data flow*



# Many standard data still unexploited ….. But many data at high frequency are informative

US example: model can run even if government shutdown (Jim Stock presentation)

| | Series | Frequency | Publication delay (in days after ref period) |
|---|---|---|---|
| 1 | Nominal Gross Domestic Product | quarterly | 28 |
| 2 | Gross Domestic Product Deflator | quarterly | 29 |
| 3 | Industrial Production Index | monthly | 14 |
| 4 | Purchasing Manager Index, Manufacturing | monthly | 3 |
| 5 | Real Disposable Personal Income | monthly | 29 |
| 6 | Unemployment Rate | monthly | 7 |
| 7 | All Employees: Total nonfarm payroll | monthly | 7 |
| 8 | Personal Consumption Expenditures | monthly | 29 |
| 9 | Housing starts | monthly | 19 |
| 10 | Single Family Homes Sales | monthly | 26 |
| 11 | Manufacturers' New Orders: Durable Goods | monthly | 27 |
| 12 | Producer Price Index: Finished Goods | monthly | 13 |
| 13 | Consumer Price Index for All Urban Consumers: All Items | monthly | 14 |
| 14 | Imports | monthly | 43 |
| 15 | Exports | monthly | 43 |
| 16 | Philadelphia survey, General Business Conditions | monthly | -10 |
| 17 | Retail and Food Services Sales | monthly | 14 |
| 18 | Conference board consumer confidence | monthly | -5 |
| 19 | Bloomberg consumer comfort Index | weekly | 4 |
| 20 | Initial Claims | weekly | 4 |
| 21 | Appliances Production Composite Index | weekly | 14 |
| 22 | Total Oil and Gas Rigs in Operation (Onshore and Offshore) | weekly | 14 |
| 23 | Coal Production Index | weekly | 14 |
| 24 | Crude Oil and Lease Condensate Production | weekly | 14 |
| 25 | Distillate Fuel Oil Production | weekly | 14 |
| 26 | Total Motor Gasoline Production | weekly | 14 |
| 27 | Kerosene-Type Jet Fuel Production | weekly | 14 |
| 28 | Residual Fuel Oil Production | weekly | 14 |
| 29 | Crushed Stone, Sand and Gravel Production Index | weekly | 14 |
| 30 | Western Softwood Lumber Production Index | weekly | 14 |
| 31 | Organic Chemicals Production Index | weekly | 14 |
| 32 | Steel Mill Products Output | weekly | 14 |
| 33 | Basic Iron and Steel Production | weekly | 14 |
| 34 | Meat Production Composite Index | weekly | 14 |
| 35 | Trucks Production | weekly | 7 |
| 36 | Autos Production | weekly | 7 |
| 37 | Electric Utilities Output | weekly | 10 |
| 38 | Total Railroad Traffic | weekly | 10 |
| 39 | Total Railroad Traffic excl. Intermodeal | weekly | 10 |
| 40 | Total Railroad Intermodal Traffic | weekly | 10 |
| 41 | Total Auto Incentives (Cash Back + Financing) | weekly | 7 |
| 42 | Total Auto Incentives (Cash Back Only) | weekly | 7 |
| 43 | Car Dealer Executives Survey | weekly | 7 |
| 44 | Autos Transactions Count | weekly | 7 |
| 45 | Baltic Dry Index | daily | 1 |
| 46 | S&P 500 Index | daily | 4 |
| 47 | Crude Oil: West Texas Intermediate (WTI) - Cushing, Oklahoma | daily | 1 |
| 48 | 10-Year Treasury Constant Maturity Rate | daily | 1 |
| 49 | 3-Month Treasury Bill: Secondary Market Rate | daily | 1 |
| 50 | Trade Weighted Exchange Index: Major Currencies | daily | 1 |

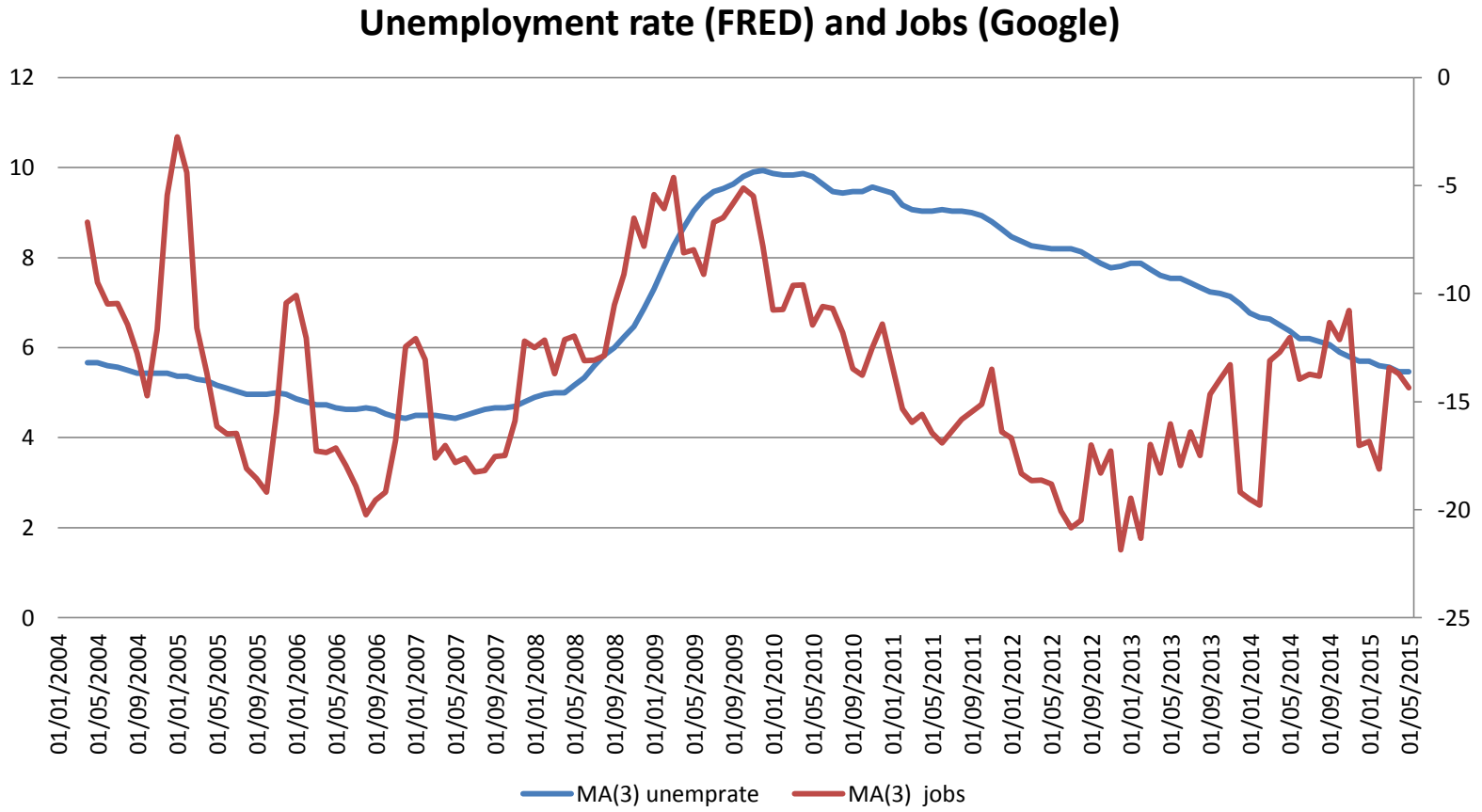
# What about google data ?

- Some evidence – mostly in sample, mostly without conditioning on standard easily available data
- Need to evaluate on the basis of a model that has all relevant details about the information problem
- Be rigorous on evaluation methods to avoid data mining (out-of sample, backcasting)
- Many standard easily available data still unexploited!

Smoothed monthly data constructed from weekly data

Unemployment (FRED); Jobs (GOOGLE)

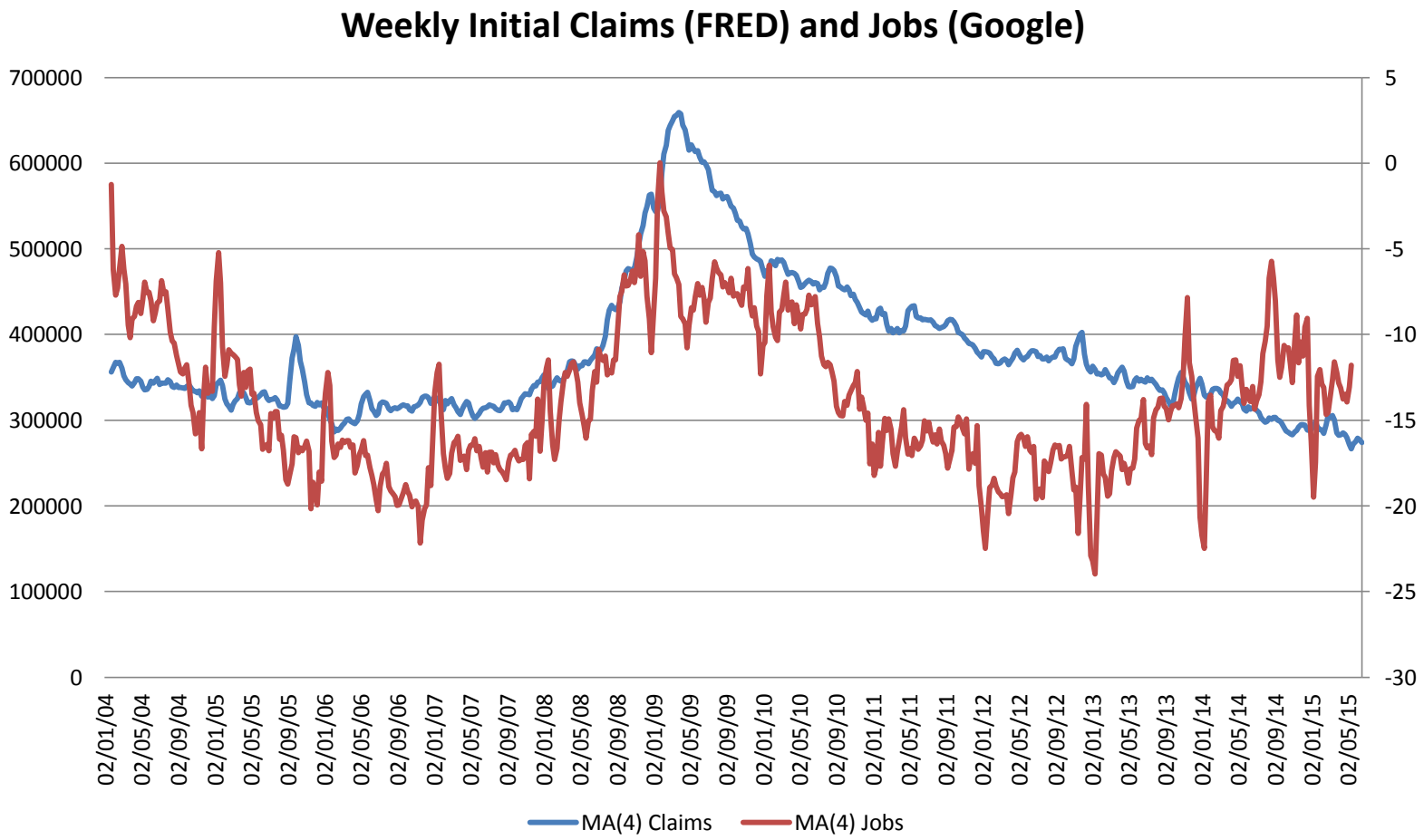
# Unemployment and google query “jobs”correlated…



Smoothed weekly data

Initial Claims (FRED); Jobs (GOOGLE)

# … but also correlated with initial claims available by standard sources



# *Conclusions (1)*

- Data are important for research and policy
- Macroeconomics had a tradition of using many data for business cycle analysis – forgotten and now revamped
- Why? More focus on empirical research; new topics: heterogeneity, information, timeliness, macro-finance
- Traditional data sources available but unexploited, US ahead of the game
- Warning! Don’t blindly import from other fields … economic data structures need research tools specifically designed

# *Conclusions (2)*

- New sources such as google potentially useful but the case has not yet been convicincely made. More research is needed but details matter …
- Methodologically need to develop models which can deal with the curse of dimensionality problem – Bayesian shrinkage is the natural way to go, need to think about identification in an environment in which data are highly correlated
- Last 20 years some useful ideas
- Lots more need to be done