Big data for economic analysis

Lucrezia Reichlin

London Business School and CEPR

London Business School
March 7, 2016

A tsunami of data

More data, standard and non standard sources, easily available, easily collected and stored

Quantifying the data deluge: the petabyte era

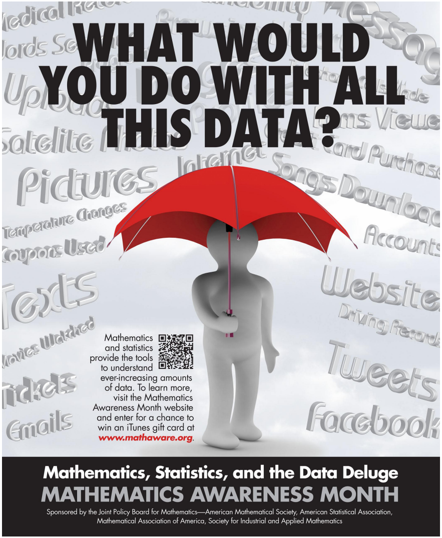
- bytes : 1 byte $\sim$ 1 letter (ascii symbol)
- kilobytes : 1 kB [1000 bytes] $\sim$ 1 page , 1 article in pdf (50-500 kB) , 1 small image
- megabytes [1 megabyte= 10^6 bytes] : 1 book
- gigabytes [1 giga= 10^9 bytes]: 1 Audio CD (700 MB), 1 DVD (5 GB), a private library
- terabytes [1 tera= 10^{12} bytes]: a public library LOC (20 TB) (digital content of U.S. Library of Congress)
- petabytes [1 peta = 10^{15} bytes]: amount of data treated by the servers of Google in one hour (1 PB)

Growth

- 90 % of the recorded data have been collected during the last two years!!!
- Most data are now digital (numbers)
(1 % en 1986, 25 % en 2000, 94% en 2007)
- In 2007, $\sim$ 300 exabytes of data stored
(61 CD-ROM per person, i.e. a stack which would go beyond the moon!)
- et $\sim$ 2 zettabytes of data exchanged! [1 zetabyte= 10^{21} bytes]

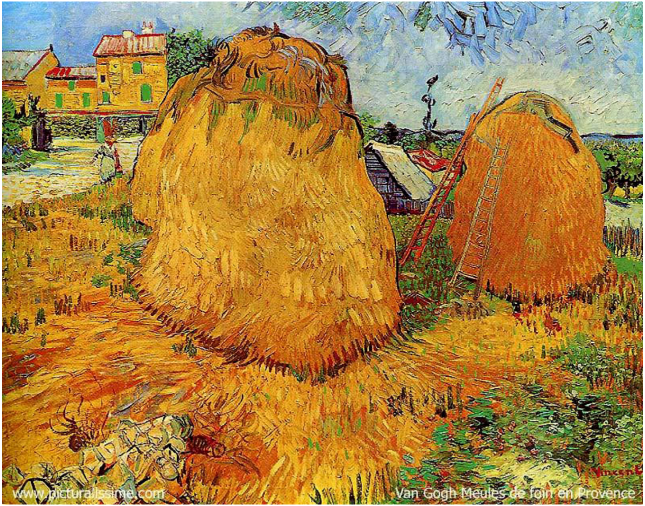
(M. Hilbert, P. López, Science 2011)

Mathematics Awareness Month April 2012



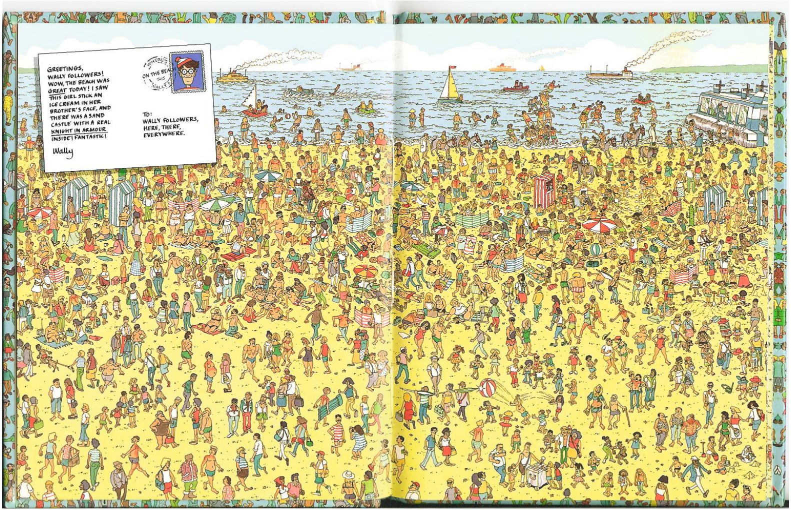
Need intelligent design to exploit them!

Finding a needle in a haystack



Can we extract a meaningful signal?

Where is Wally? (Martin Handford)



Model for automatic detection of people from complex systems (computer vision)

Big challenges ...

for

- mathematicians
- statisticians
- computer scientists, engineers, etc.

in order to develop automatic procedures for extracting useful information from huge amounts of data.

$\rightarrow$ rapid development of (new) research fields:

- Computer vision
- Data Mining
- Statistical Learning ("Machine Learning")
- Bioinformatics, etc.

... in all scientific areas

- Physics
- Astronomy
- Geophysics
- Biology (chiefly genomics, proteomics)
- ... and also economics, finance and social sciences

What about economics?

- New statistical methods for estimating models with large data
- But also use of new data: social media, google queries, geo-locational, ...

Many recent and less recent examples: tick-by-tick data, portfolio optimization, forecasting and stress tests, health data, ...

Most of these applications involve *large* data rather than *big* data but exploit new statistical models to deal with the curse of dimensionality problem: dimension reduction, sparsity, compression, new approaches to algorithms, ...

Is this a revolution for the field?

My answer in a nutshell: not clear yet!

- Possibly a change in methodological philosophy: more emphasis on prediction rather than on causal relations
- Emphasis on real time analysis
- Democratization of statistics

$\Rightarrow$ Not the end of theory but useful fresh air

Discussion on selected issues

1. The curse of dimensionality and machine learning

   what is this all about?

2. What works with macro data?

   The curse and blessing of collinearity

3. Google data and real time now-casting

   Not very useful

4. Prediction and causality

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?

$\Rightarrow$ The sample variance is inversely proportional to the degrees of freedom (sample size minus number of parameters)

$\Rightarrow$ When number of parameters becomes large, the degree of freedoms go to zero or become negative and the precision of the estimates deteriorate

This is the curse of dimensionality

Solutions which have been used in econometrics

- *Factor analysis*:
  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), many applications in macro, theory and empirics for large dynamic models
- *Principal components*:
  extract first PCs / if few factors drive the dynamics of the data it works well
- *Penalized regression*:
  limit estimation uncertainty via shrinkage [machine-learning]

These methods are related: either aggregate variables or select or both

Problems with traditional approach

A simple illustration:

Forecast y_t using a **large** information set:

$$\hat{y}_{T+h|T} = \hat{\beta}' X_T$$

Estimate $\hat{\beta}$ via OLS, i.e. maximize the in-sample fit of the model:

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=1}^{T-h} \left(y_{t+h} - \beta' X_t \right)^2$$

$$\Longrightarrow \boxed{\hat{\beta} = (X'X)^{-1} X'y} \Rightarrow \boxed{\hat{y}^{OLS}_{T+h|T} = \hat{\beta}' X_T}$$

Problem!! If the size information set (n) is too large relative to the sample size (T) then OLS forecasts are poor or unfeasible: <u>curse of dimensionality</u>.

A cure for the illness: Penalized regression

To stabilize the solution (estimator), use extra constraints on the solution or, alternatively, add a penalty term to the least-squares loss

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

Example: ridge: penalized regression (L_2 norm)

$$\hat{\beta}^{ridge} = \arg\min_{\beta} \sum_{t=1}^{T-h} \left(y_{t+h} - \beta' X_t\right)^2 + \nu \sum_{i=1}^{n(p+1)} \beta_i^2$$

$$\boxed{\hat{\beta}^{ridge} = \left(X'X + \nu I\right)^{-1} X'y}$$

• Ridge is a form of linear '*shrinkage*', where the components of $\hat{\beta}_{ols}$ are shrunk uniformly towards zero

• it is a kind of "regularization" which provides the necessary dimension reduction and increases the bias to decrease the variance

Bayesian language

- Penalized regression can be reinterpreted as a Bayesian regression

limit length β + estimate coefficients as the posterior mode to compute forecast

or ... shrink regression coefficients to zero via priors

DATA (complex/rich) + PRIOR (naive/parsimoniuos)

In the case of the example: i.i.d. prior on β: $\Phi_0 = \sigma_\beta^2 \boldsymbol{I}$

Several ways of doing it

Two extreme choices:

- Normal prior - give a weight to all regressors (eg ridge) / similar to PCs give more weight to large sources of variations
- Double exponential - allows for variable selection by enforcing sparsity, *i.e.*, the presence of zeroes in the vector β of the regression coefficients (also known as '*Lasso regression*')

Sparsity?

"Entia non sunt multiplicanda sine necessitate"

William of Ockham ($\sim$ 1288 - 1348)

Macro problems: does it matter the way we do it?

- Macro and financial data are highly collinear: few factors (shocks) explain the bulk of dynamics.
- As a consequence variable selection or aggregation methods deliver the same results [empirics and theory]

$\star$ Collinearity is curse: variable selection gives unstable results

$\star$ But is is also a blessing: All methods allow to capture large sources of variations

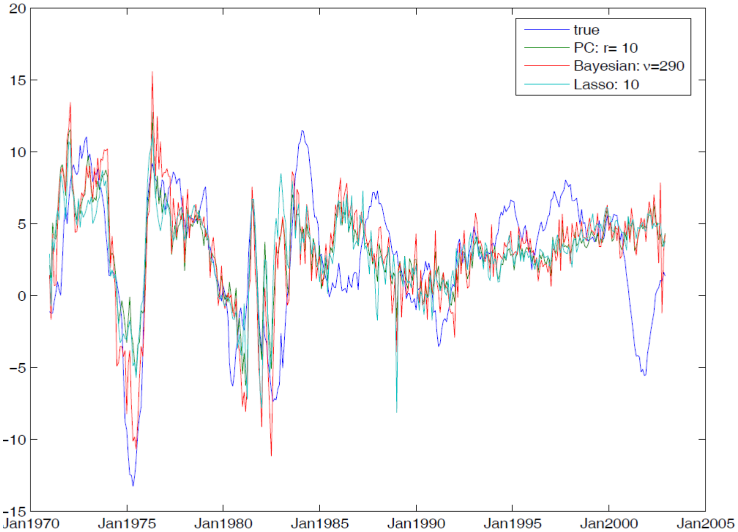
A classic result: two shocks drive the business cycle

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

Forecasting industrial production: PCs, ridge and Lasso - 200 variables



The end of Theory?

The macro approach: a compromise

- Extract shocks from large data and their lags: combinations of prediction errors
- Identify structural shocks using minimal theory
- Do not identify all coefficients but compute impulse response functions to structural shocks
- Many useful applications: the effect of unexpected policy changes, stress tests and conditional forecasting for economic analysis

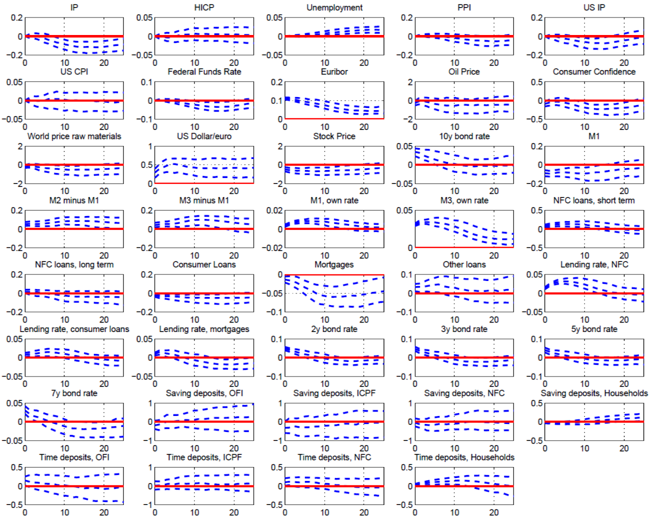
Large Bayesian VAR

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

- By shrinking in relation to the sample size can estimate with hundreds of variables
- This avoids over-fitting
- Many useful applications which were used for small models can now be used for large models

Blessing of dimensionality: in large models if correlations are stable, density forecasts work well [narrow bands]. True for both unconditional and conditional forecasts

Large Bayesian VAR: the effect of the monetary policy shock



Large Bayesian VAR: stress tests

- Macro risk is characterized by correlations: banking system overly exposed to risk in the up and too risk adverse in the down: need to look at aggregate risk
- Combine macro variables and balance sheet variables
- Construct stress scenarios by conditioning on specific assumptions

Other applications in macro: now-casting

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!

Following the calendar

Conjuctural information: This Week

Jun 17 - Jun 23									Filter On
Date	10:07am	Currency	Impact		Detail	Actual	Forecast	Previous	Chart
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	
Fri Jun 22	6:30pm	USD		FOMC Member Pianalto Speaks					

Following the calendar

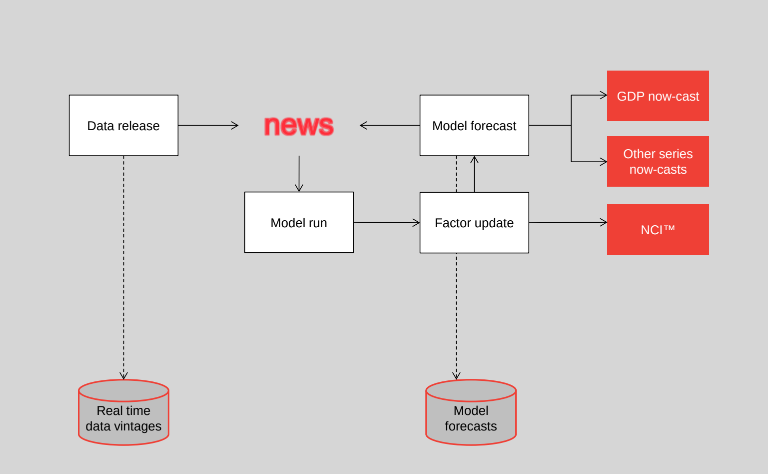
Conjuctural information: Next Week

Jun 24 - Jun 30 — Filter On

Date	10:08am	Currency	Impact		Detail	Actual	Forecast	Previous	Chart
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	
	3:55pm	USD		Revised UoM Inflation Expectations				3.0%	

The Now-Casting platform

NOW-CASTING.COM
economics in real time

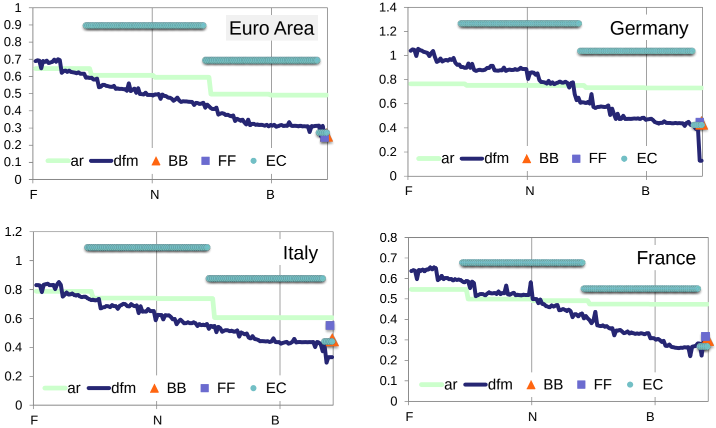


What have we learned in years of experience running an automatic procedure with no judgement?

- 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

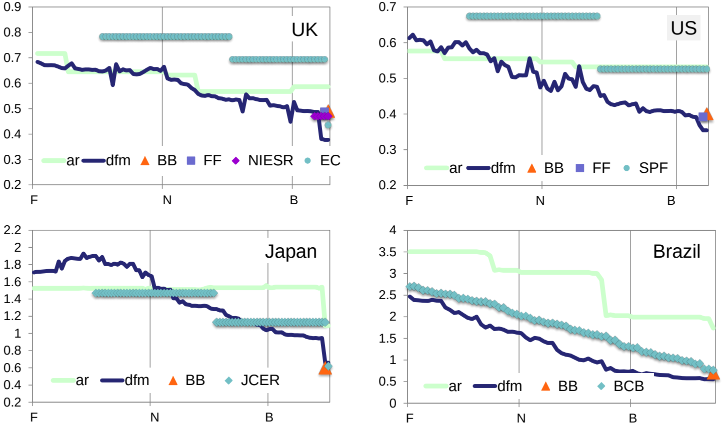
Data help

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



Data help

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



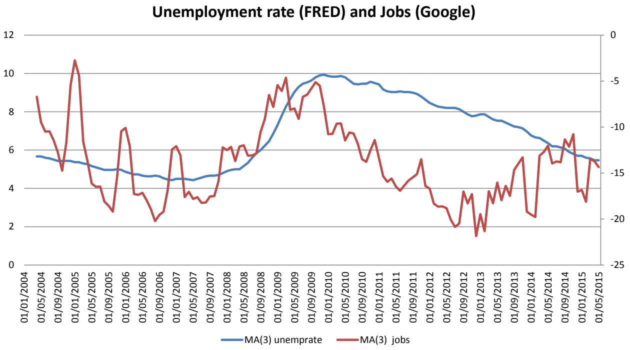
Data are available but often unexploited: the US government shutdown

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 Sorftwood 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

Do non standard timely data help? Unemployment and google query jobs are correlated

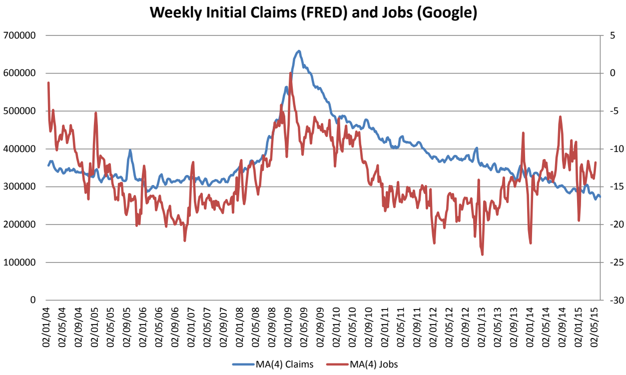
Smoothed monthly data constructed from weekly data
Unemployment (FRED); Jobs (GOOGLE)



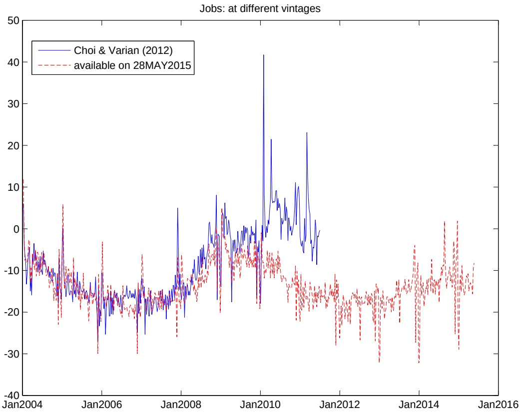
but also correlated with initial claims available by standard sources

Smoothed weekly data

Initial Claims (FRED); Jobs (GOOGLE)



and sampling errors and changes in the algorithm lead to instability and lack of robustness



Some examples beyond macro

- *Health economics*: prediction of whether replacement surgery for patients with osteoarthritis will be beneficial for a given patient or not, based on more than 3000 variables recorded for about 100 000 patients.
- *Economics and law*: machine-learning algorithms can be more efficient than a judge in deciding who has to be released or go to jail while waiting for trial because of danger of committing a crime in the meanwhile
- *Microeconometrics*: controlling for many covariates in order to better identify treatment effect, many instruments, ...
- *All fields*: combining models for robustness (empirical growth literature: i run 1 million regressions!)

Sherlock Holmes

“Data! Data! Data!” he cried impatiently. “I can’t make bricks without clay”
(Arthur Conan Doyle, The Adventure of the Copper Beeches, 1892)

“It is a capital mistake to theorize before one has data. Insensibly one begins to twist facts to suit theories, instead of theories to suit facts.”
(Arthur Conan Doyle, A Scandal in Bohemia, 1892)