Big Data, Digital Industrial Revolution and Implications on Economist Profession

Manh Nguyen, Cafe-Seminaire

May 20, 2017

Introduction

Introduction

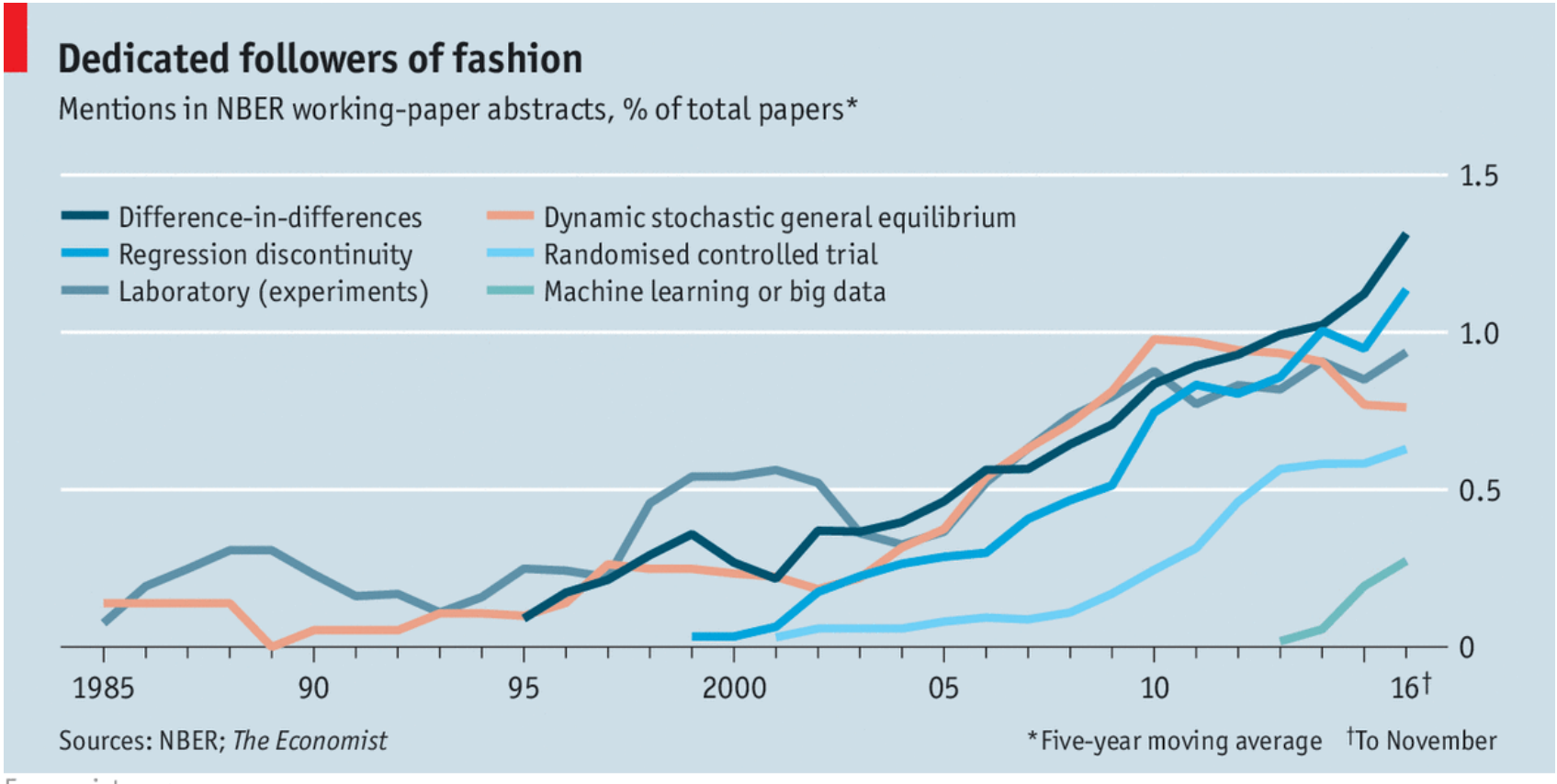

Figure 1: Economic Model Trends

Machine Learning versus Statistics

Statistical criteria: AIC, BIC and goodness of fit

Suppose that we have a statistical model M of some data. Let $\hat{L}$ be the maximized value of the likelihood function for the model, let k be the number of estimated parameters in the model and n be the training dataset:

$$AIC = 2k - 2ln(\hat{L}) \tag{1}$$

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1} \tag{2}$$

$$BIC = ln(n)k - 2ln(\hat{L}) \tag{3}$$

And other goodness of fit: R^2, chi-squared, Shapiro, t-test, etc.

ALL ARE COMPUTED ON TRAINING DATASET

Machine Learning approach: Cross-Validation

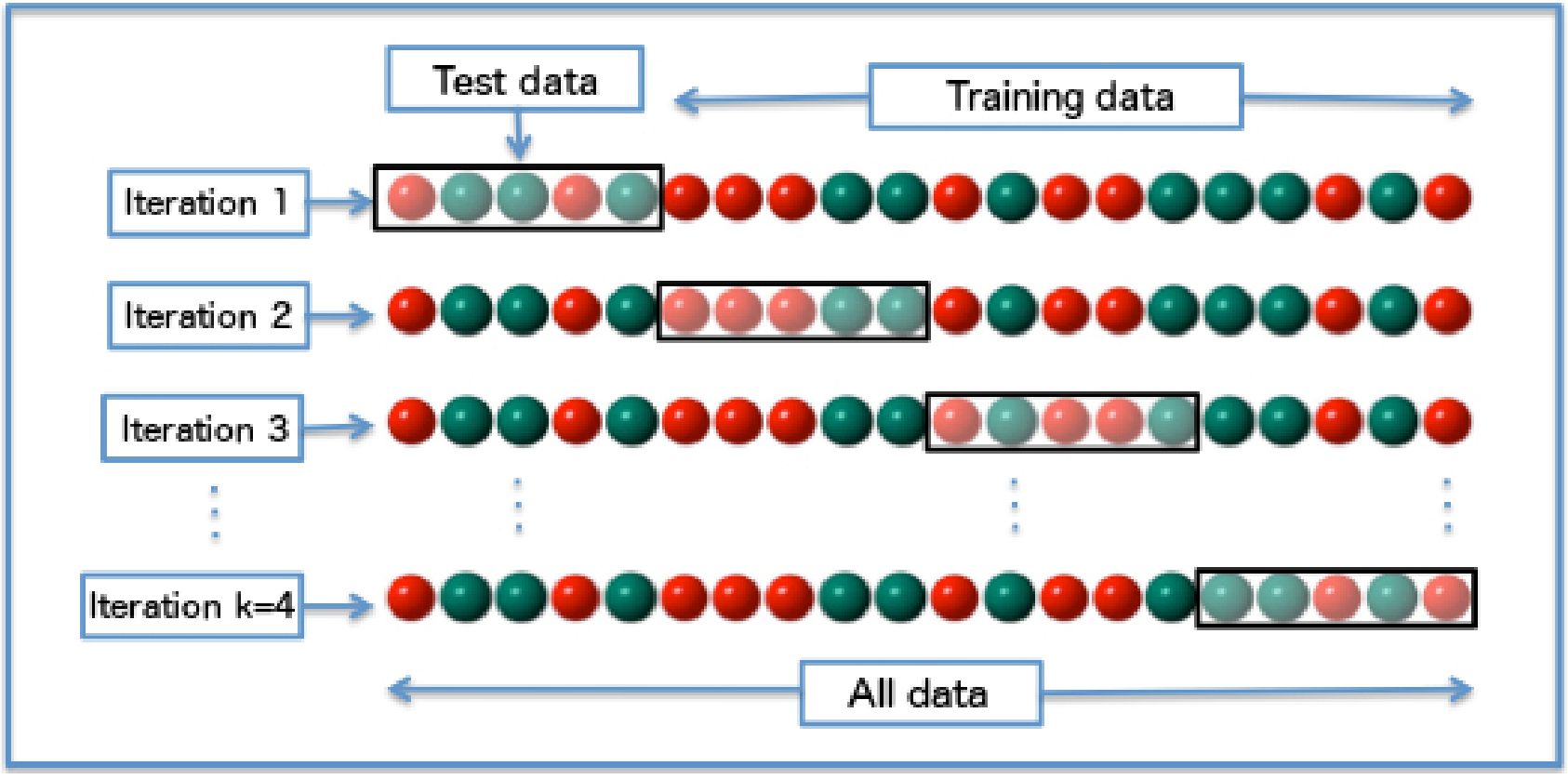

Figure 2: 4 fold cross-validation

Model complexity versus Model generalization

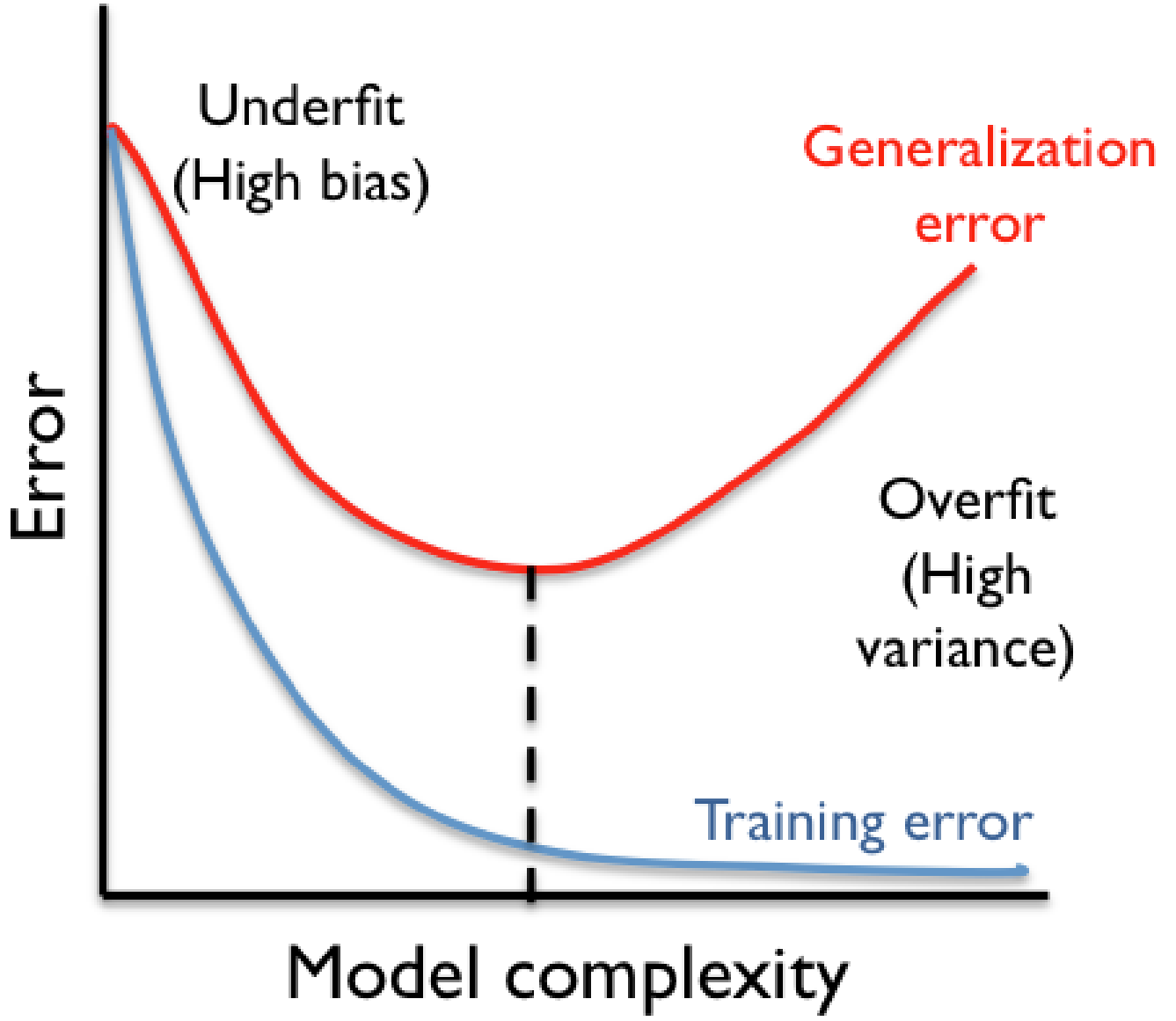

Figure 3: Model complexity versus Model generalization

To explain or to predict ? (1)

Galit Shmueli (2010) "To explain or to predict ?":

- The AIC is better suited to model selection for prediction as it is asymptotically equivalent to leave-one-out cross-validation in regression, or one-step-cross-validation in time series. On the other hand, it might be argued that the BIC is better suited to model selection for explanation, as it is consistent.
- P-values are associated with explanation, not prediction. It makes little sense to use p-values to determine the variables in a model that is being used for prediction.
- Multicollinearity has a very different impact if your goal is prediction from when your goal is estimation. When predicting, multicollinearity is not really a problem provided the values of your predictors lie within the hyper-region of the predictors used when estimating the model.

To explain or to predict ? (2)

- An ARIMA model has no explanatory use, but is great at short-term prediction.
- How to handle missing values in regression is different in a predictive context compared to an explanatory context. For example, when building an explanatory model, we could just use all the data for which we have complete observations (assuming there is no systematic nature to the missingness). But when predicting, you need to be able to predict using whatever data you have. So you might have to build several models, with different numbers of predictors, to allow for different variables being missing.
- Many statistics and econometrics textbooks fail to observe these distinctions. In fact, a lot of statisticians and econometricians are trained only in the explanation paradigm, with prediction an afterthought. That is unfortunate as most applied work these days requires predictive modelling, rather than explanatory modelling.

Model oriented or data oriented ?

Model dominates data:

- Small dataset, little variables (features), backed by theoretical foundation
- Suppose that the phenomenon follows a model or some model with null hypothesis, then test whether we can reject the null hypothesis.

Data dominate model:

- Big data, enormous features, no theoretical foundation yet due to complexity of problem
- Suppose that the problem is huge, we only seek to predict what will turn out NEXT. Test whatever it makes sense due to cross validation criteria.

But, it is really . . .

Model and data domination go back and forth, go from A to B, from B to C. Ideally, model and data would always meet. Start from model approach, repeat for all models possile. Start with data approach, repeats to find good models.

Like theories and use case.

Theories/Models are lightweight, easilly to transfer and spill over. Data are more specific, heavy.

No winner or loser but that is the gist of it.

Right now, it is just the explosion of data (Big Data). Algorithms are just turning around data.

Big Data and The Fourth Industrial Revolution

Big Data

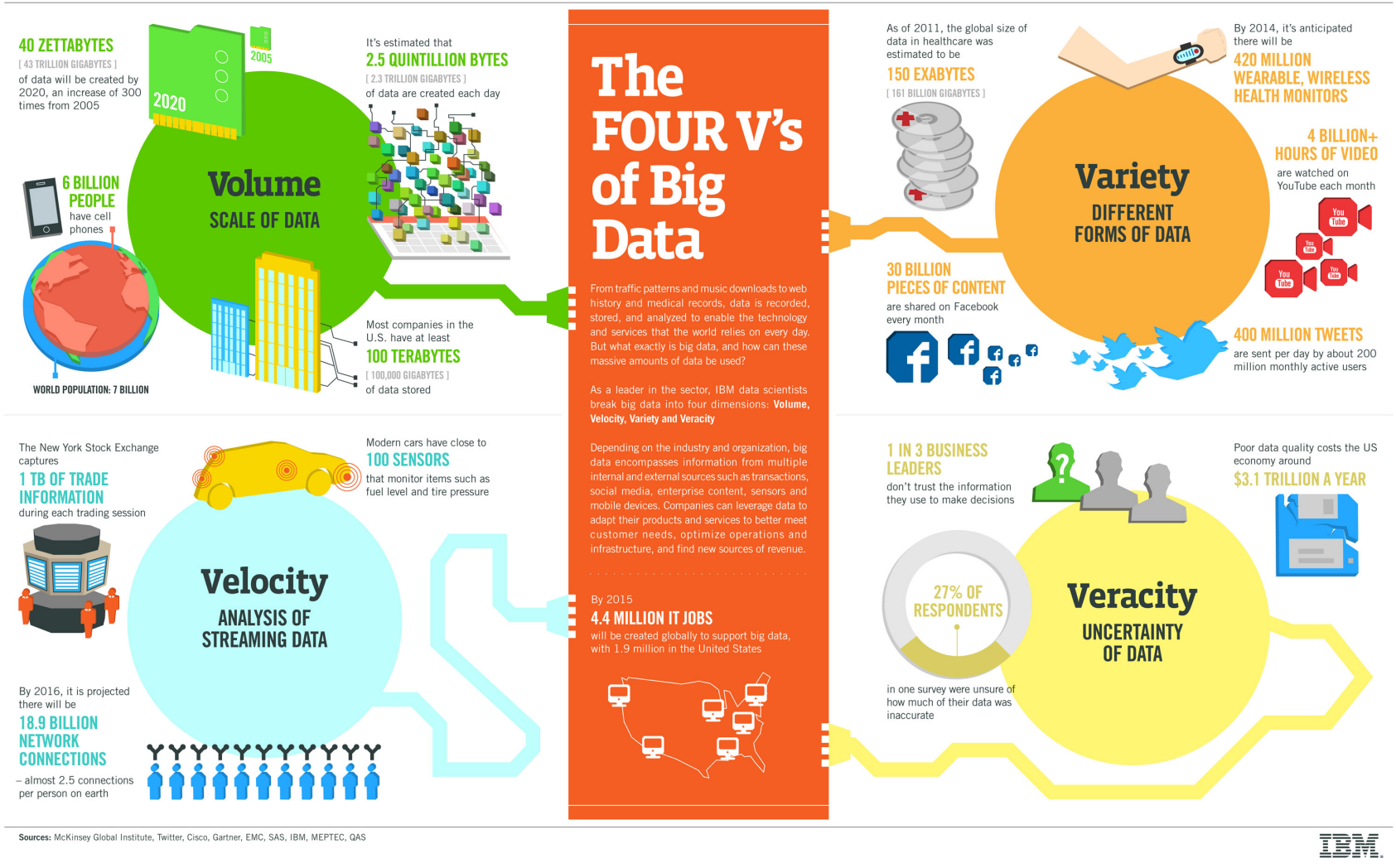

Figure 4: The 4 Vs of Big Data

The fourth industrial revolution

- 1st industrital revolution: 1760 - 1820 or 1840: steam engine, iron and textile industries
- 2nd industrial revolution: 1870 - 1914: steel, oil, electricity, electric power to create mass production
- 3rd industrial revolution or digital revolution: 1980s - present: PCs, the Internet and IT technologies
- 4th industrial revolution: 2004 - ongoing the big data intelligence: robotics, artificial intelligence, nanotechnology, biotechnology, IoTs, 3D printing, autonomous vehicules

Big data infrastructures have just reached mature pharse. It is now big data analytics: how to analyse these data.

What is news in computing ?

Scalability:

- Cluster computing: data are distributed in NoSQL database
- Scalable algorithms over distributed database
- Parallel computing over GPUs for deep learning (1000s of cores)

No one super computer but a super cluster of computers.

Implications

Impications for Economic Studies

- The digital economy: more and more digitally intergrated. Jobs, regulations, unemployment.
- Much more other measures for economics: trends, sentiment analysis, etc over the Internet versus existing statistic measures.
- Spatial data and the gravity equation literature in trade theories
- Deep learning or NLP helps to trace back historical texts, analyse the whole human knowledge. These technologies are available right now. Just make use of that.
- Not only text, but images and videos: we have time series of number, how about time series of more sophisticated objects, such as a human face, traits of characters etc
- Agent-based model: possibility to test through data at individual level (not just firm level)
- Surveys still exist but are still costly, versus other digital survey (larger, cheaper but more difficult to deal with)
- etc ...

Implications for Vietnam ?

We did not know much about the 3 previous industrial revolutions. Now, there are a lot of people who know and are deep in this revolution.

The Road Map to the Big Data Era

The Road Map

- Learn a programming language for:
  - Data querrying
  - Data cleansing
  - Algorithm tuning for data
  - Reporting results
- Scripting languages: R, Python, Perl, Javascript are so easy to learn
- Learning is never easier: MOOCs (2013 the year of MOOCs) coursera, edx, audacity etc
- Codes are basically open source. Everything is on GitHub.
- For economic studies: learn R. SAS is still good but it is not free.

But still, one always should start with a bunch of questions.

Referecences

References

- https://www.economist.com/news/finance-and-economics/21710800-big-data-have-led-latest-craze-economic-resea
-