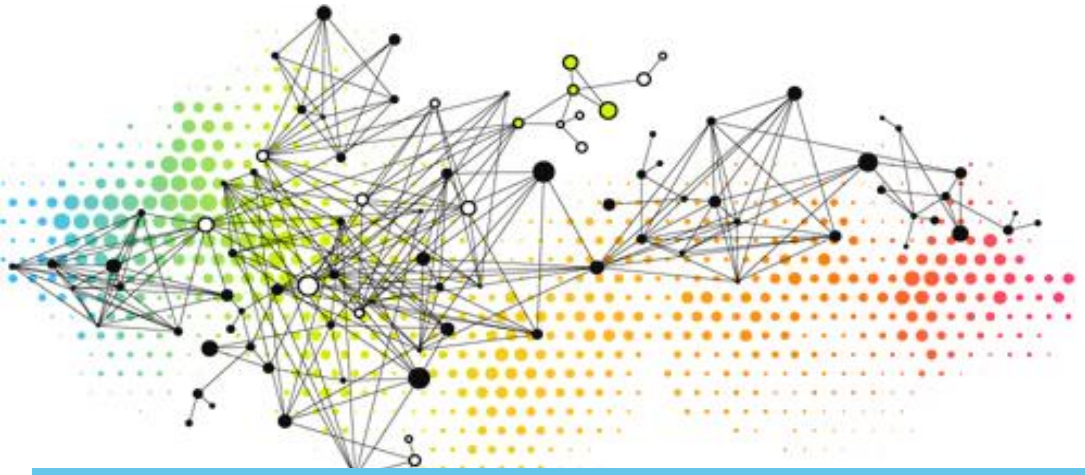


Statistics Department

Big Data: Potential, Challenges and Statistical Implications

Gabriel Quirós

Deputy Director, STA, IMF

Economic and Financial Regulation in the Era of Big Data

Banque de France

November 24, 2017

Outline

I. Background

II. What does big data mean?

III. Potential

IV. Statistical implications and potential

V. Challenges

VI. Dos and Don’ts of Big Data for statistics

I. Background

- **STA set up an internal group to develop a vision on big data for statistics:**
  - *potential* of big data to benefit macro-economic and financial statistics?
  - organizational, budgetary, and, in particular, methodological *challenges* that come with incorporating big data?
  - and strategic *statistical implications* for national and international organizations moving forward

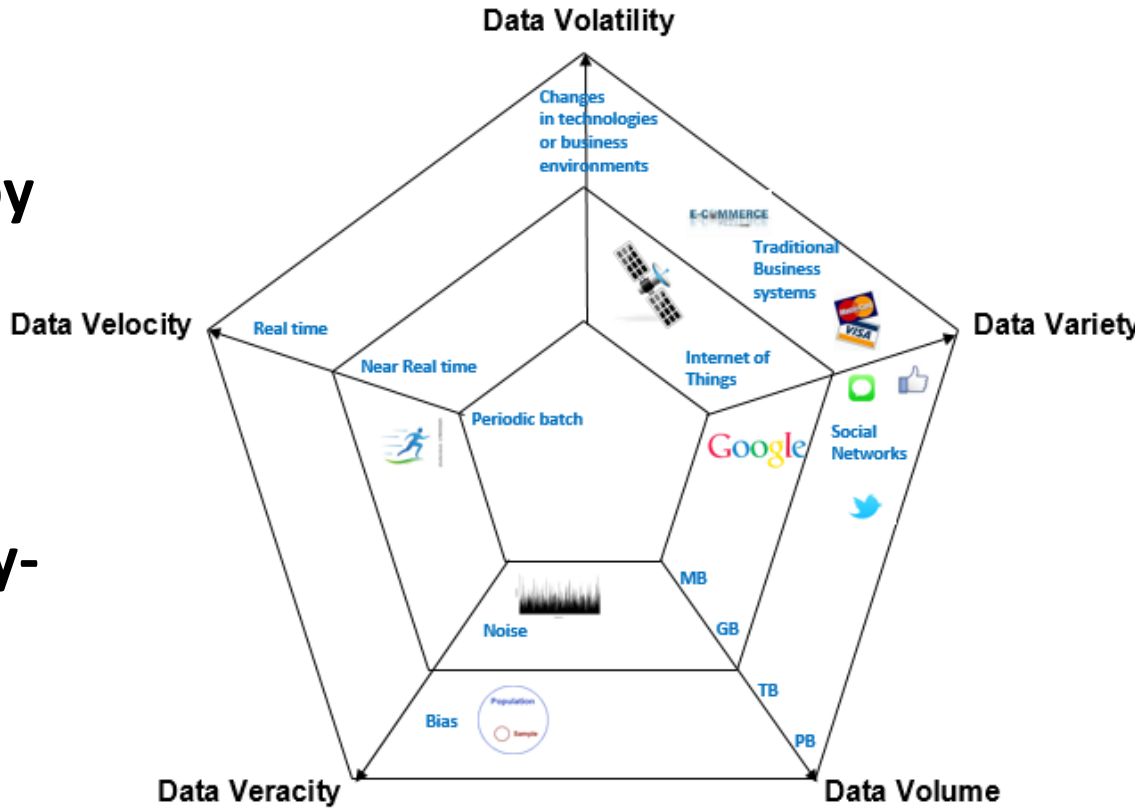
I. Background

- **Some key points to consider:**
  - big data are not static, but a far-ranging *evolving concept* that requires a long-term vision
  - a strategic organizational plan to deliver measurable and high-scale results trumps individual and scattered applications of big data at national/international organizations
  - this discussion is a *starting point for further research* and detailed analyses on the use of big data to directly and indirectly support IMF surveillance work

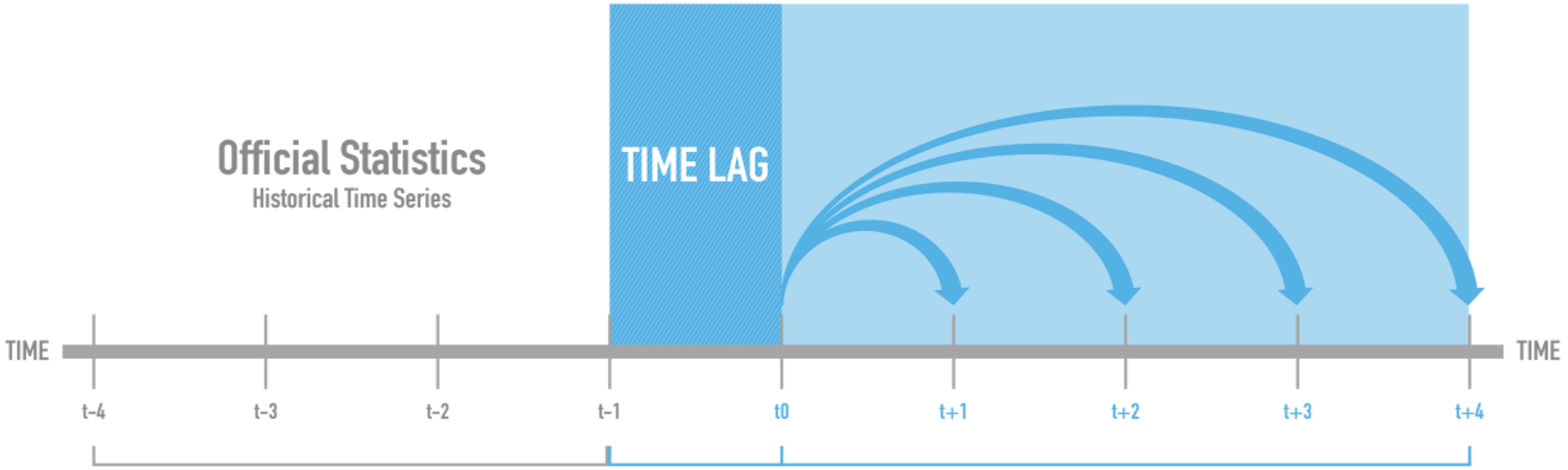
II. What does big data mean?

No straight-forward definition

- Doug Laney, 2001
- **Big data characterized by the "5Vs"**
  - High-volume, high-velocity, high-variety
  - Veracity and volatility
- **Big data ("found") as 'by-products'**
  - Social networks
  - Business operations
  - The Internet of Things



III. Where, how is the Potential of Big data for statistics?



IV. What Challenges come with Big Data?

- **Data Quality**
  - quality assessments of indicators will be crucial to minimize governance, political, and reputational risks
  - statistical techniques and methodologies best practices are needed to specifically address *veracity* and *volatility*
  - big data for uncovering meaningful insights, trends, and sentiments may underlie different quality assessments compared to using big data in official statistics
    - continuation of consistent and harmonized historical time series is still needed
  - metadata are key to assess and interpret new data sources

IV. What Challenges come with Big Data?

Data Access

- Proprietary data held by the private sector
  - Public-Private Partnerships that safeguard independence, privacy and confidentiality
- Data that companies own may evolve from a byproduct to becoming a major asset
- Regular licensing costs come in addition to substantial investments into processing and storage solutions
- Risk of volatility persists
- Best practices for building lasting relationships between data owners and data users are needed (UN Global Working Group)

IV. What Challenges come with Big Data?

- **New Skill Profiles**
  - Special career stream for data scientists
- **Multi-disciplinary project teams needed to make big data speak**
  - Experts from different professional backgrounds work together
- **IT implications**
  - Sharing of software codes and algorithms; open-source software; cloud-computing

V. Implications for Statistical Domains and their potential (1/2)

Data Origin+	Data Type	Data Source and Techniques	Potential Indicators Derived	Statistical Domains	What May be the Potential?*
Social Networks	Social networks, blogs and comments 1100. Social Networks: Facebook, Twitter, LinkedIn 1200. Blogs and comments 1600. Internet searches on search engines (Google) 1700. Mobile data content: text messages, Call Detail Record, Data Detail Record, Location update, Radio coverage updates Online news	**Google** trends and search data	now-cast GDP now-cast unemployment consumer sentiment car and property sales	National accounts External sector statistics Financial Statistics Price statistics	**2**
		Mobile phone system data (electronic money schemes, e.g. M-Pesa) Peer-to-peer transactions	financial inclusion indicators remittances, regional disposable income, consumption patterns poverty reduction SDG "Gender Equality" economic growth	National accounts External sector statistics Financial Statistics Price statistics	**1,3**
		Twitter tweets	consumer confidence index border mobility, tourism, transitioning of migrants now-cast food prices sentiment and topic trend analysis	Mobility and urban statistics Price statistics Demographic and social statistics	**1,2,3**
		Web-scraping of **Facebook posts, Wikipedia articles**	geopolitical risk indicators price changes civil protests/labor strikes and national security events consumer sentiment inclusive infrastructure for sustainable development	Price statistics National accounts Demographic and social statistics Labor Statistics	**1,2**
		Call Detail Record data	SDGs indicators, travel/tourism, transport, migration	Mobility and urban statistics	**1,3**
Traditional Business Systems	**Data produced by public agencies** Administrative data	**Taxation registers**	consumer spending small business' income nonresident businesses controlled by resident parent corporations business profiling flight reservation system	National accounts Price Statistics External sector statistics Labor statistics Tourism statistics Transportation statistics	**2, 3**
		Population/business registers	multi-sourcing to derive population and housing census population structure	National accounts Demographic and Social Statistics	**3**
	Data produced by businesses 2210. Commercial transactions	**SWIFT data** on transaction quantities and financial market prices	global financial flows network concentration cross-border transactions export/import indicators	National accounts Price statistics External sector statistics Financial Statistics	**2,3**

V. Implications for Statistical Domains and their potential (2/2)

Data Origin+	Data Type	Data Source and Techniques	Potential Indicators Derived	Statistical Domains	What May be the Potential?*
	2220. Banking/stock records 2230. E-commerce 2240. Credit cards Business websites Scanner data		withdrawal of correspondent banking relationships trade financing		
		Web-scraping to collect price data from online retailers	daily inflation turning points in inflationary trends e-commerce index	Price statistics Financial statistics	2,3
		Web-scraping business websites	enterprise profiling job vacancies	National accounts Financial statistics Labor statistics	2,3
		Scanner data Prices and quantities	national and regional consumer prices household income and expenditure	Price statistics National accounts Financial statistics	2,3
		Credit card data	consumer spending growth trends of the retail sales	National accounts External sector statistics	
Internet of Things (machine-generated data)	**Data from sensors** **311. Fixed sensors** 3111. Home automation 3112. Weather/pollution sensors 3113. Traffic sensors/webcam 3114. Scientific sensors **312. Mobile sensors (tracking)** 3121. Mobile phone location 3122. Cars 3123. Satellite images	GPS positioning/tracking data	travel services exports/imports trip duration inbound/outbound international travelers remoteness index traffic intensity	National accounts External sector statistics Demographic statistics Transport statistics Urban statistics Tourism statistics Population statistics	1,2,3
		Traffic/Road sensors	proxy of economic growth/health commuting time traffic intensity incoming/outgoing traffic travel/tourism	National accounts External sector statistics Transport statistics Tourism statistics Mobility statistics	1,2,3
		Satellite imagery Research and mapping of weather and climate data	improved geographical localization of statistical units and assets spatial sampling frame for output measurement land use and geostatistical cartography crop planting area, land use and agricultural output population and asset location as proxy for SDG "Gender Equality"	National accounts Price statistics External sector statistics Demographic and social statistics Transport statistics Agricultural statistics Demographic and urban statistics	1,3
		Smart meters (energy consumption measures)	non-occupancy rates household consumption electricity supply and consumption price differentials household structure and size	Environmental and Energy statistics National Accounts Price statistics Demographic and Social Statistics Transportation Statistics Geo-Spatial Statistics Agricultural Statistics Rural and Population Statistics	1,2,3

VI. Dos and Don'ts of Big Data

- In connection to the respective statistical domains, a number of Dos and Don'ts from big data can be identified, which are unevenly distributed across statistical domains
- The Dos and Don'ts are largely driven by the essence of big data: by-products of private technological and business models that capture behavior of consumers, corporates, banks, individuals or government agencies
- Big data are particularly promising to enhance directly or indirectly statistics on transactions, less so on stocks

VI. Dos and don'ts of big data

Dos

Big data, particularly promising at helping measure:

- ✓ "soft" information: sentiment, alerts, reactions…
- ✓ consumer behavior and patterns (e.g. Amazon, Google searches and 'clicks', social networks,…)
- ✓ Tourism (e.g. roaming information, Google searches, credit cards, click-stream data, …)
- ✓ Financial flows (e.g. SWIFT, mobile phones, …)
- ✓ Prices (scanner data,…)
- ✓ Job vacancies and labor skills (e.g. LinkedIn,…)
- ✓ big data provides granular, microdata
- ✓ …..

VII. Dos and Don'ts of Big Data

Don'ts

- Sample representativeness: bias towards more modern and dynamic economic activities and social behavior
- Big data less suited for stocks, i.e. total financial assets and liabilities of firms, households, government, non residents, both at micro and macro levels
- Revaluation and other volume changes, particularly important in monetary and financial statistics
- As by-product, long time-series based on big data are inexistent and will be fragile because instability from business and technological changes, discontinuity in data provision
- Privacy and confidentiality of personal, firm-level data