Big Data for Central Banking

G. Sivakumar
Computer Science and Engineering
Bharatiya Vidyalaya and University of Mumbai (IIT Bombay)
siva@iitb.ac.in

- Data Scientist’s Dream: AI/ML meets Web 3.0 & SMAC + IoT
- Data Engineer’s Job: Big Data Frameworks- Hadoop, Spark, Flink
- Data Governance’s Nightmare: Quality, Security, Privacy, RoI
- RBI’s CIMS (Centralized Information Management System)
Note: The risks of analytical thinking and fragmentation of knowledge

- The New York Fed Staff Nowcast for 2018:Q3 stands at 2.0%.
- News from this week’s data releases decreased the nowcast for 2018:Q3 by 0.4 percentage point.
- The negative impact of higher than expected manufacturers’ inventories drove most of the decrease.

### Data Flow (Aug 24, 2018)

<table>
<thead>
<tr>
<th>Model Update</th>
<th>Release Date</th>
<th>Data Series</th>
<th>Actual</th>
<th>Impact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aug 24</td>
<td>8:30AM Aug 24</td>
<td>Manufacturers’ inventories: Durable goods</td>
<td>1.25</td>
<td>-0.32</td>
</tr>
<tr>
<td>Aug 24</td>
<td>8:30AM Aug 24</td>
<td>Manufacturers’ unfilled orders: All industries</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Aug 24</td>
<td>8:30AM Aug 24</td>
<td>Manufacturer’s shipments: Durable goods</td>
<td>-0.21</td>
<td>-0.02</td>
</tr>
<tr>
<td>Aug 24</td>
<td>8:30AM Aug 24</td>
<td>Manufacturers’ new orders: Durable goods</td>
<td>-1.69</td>
<td>-0.05</td>
</tr>
<tr>
<td>Aug 25</td>
<td>10:00AM Aug 25</td>
<td>New single-family houses sold</td>
<td>-1.72</td>
<td>-0.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data revisions</td>
<td>-0.01</td>
<td></td>
</tr>
</tbody>
</table>

Aug 23: 2.39
STAFF REPORTS

Macroeconomic Nowcasting and Forecasting with Big Data

November 2017 Number 830

JEL classification: C32, C53, C55, E32

Authors: Brandyn Bok, Daniele Caratelli, Domenico Giannone, Argia M. Sbordone, and Andrea Tambalotti

Data, data, data . . . Economists know it well, especially when it comes to monitoring macroeconomic conditions—the basis for making informed economic and policy decisions. Handling large and complex data sets was a challenge that macroeconomists engaged in real-time analysis faced long before “big data” became pervasive in other disciplines. We review how methods for tracking economic conditions using big data have evolved over time and explain how econometric techniques have advanced to mimic and automate the best practices of forecasters on trading desks, at central banks, and in other market-monitoring roles. We present in detail the methodology underlying the New York Fed Staff Nowcast, which employs these innovative techniques to produce early estimates of GDP growth, synthesizing a wide range of macroeconomic data as they become available.

Available only in PDF
Evolution of Atlanta Fed GDPNow real GDP estimate for 2018: Q3
Quarterly percent change (SAAR)

Sources: Blue Chip Economic Indicators and Blue Chip Financial Forecasts
Note: The top (bottom) 10 forecast is an average of the highest (lowest) 10 forecasts in the Blue Chip survey.
Changes in subcomponent contributions to GDPNow real GDP growth forecasts

- Consumer spending (PCE)
- Residential investment
- Net exports
- Nonresidential fixed investment
- Change in private inventories
- Government spending

Total change in GDP forecast

Source: Atlanta Fed

FEDERAL RESERVE BANK OF ATLANTA

Big Data for Central Banking
## Atlanta Fed GDPNow estimates for 2018: Q2. growth rates and changes

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>30-Apr</td>
<td>Initial nowcast</td>
<td>4.1</td>
<td>3.3</td>
<td>9.1</td>
<td>4.4</td>
<td>2.9</td>
<td>3.5</td>
<td>0.5</td>
<td>6.1</td>
<td>3.6</td>
<td>7</td>
<td>26</td>
</tr>
<tr>
<td>31-May</td>
<td>Adv. Econ. Indicators/GDP (5/30), Personal Inc., NIPA underlying detail</td>
<td>4.7</td>
<td>3.4</td>
<td>5.4</td>
<td>6.4</td>
<td>5.6</td>
<td>0.8</td>
<td>0.6</td>
<td>7.3</td>
<td>2.9</td>
<td>19</td>
<td>43</td>
</tr>
<tr>
<td>4-Jun</td>
<td>M3 Manufacturing, Auto sales</td>
<td>4.5</td>
<td>3.3</td>
<td>4.3</td>
<td>6.4</td>
<td>7.2</td>
<td>1.8</td>
<td>0.4</td>
<td>7.5</td>
<td>3.2</td>
<td>18</td>
<td>42</td>
</tr>
<tr>
<td>5-Jun</td>
<td>ISM Nonmanufacturing Index</td>
<td>4.6</td>
<td>3.3</td>
<td>4.4</td>
<td>6.4</td>
<td>7.2</td>
<td>1.9</td>
<td>0.4</td>
<td>7.5</td>
<td>3.2</td>
<td>18</td>
<td>42</td>
</tr>
<tr>
<td>6-Jun</td>
<td>International trade</td>
<td>4.5</td>
<td>3.4</td>
<td>4.0</td>
<td>6.4</td>
<td>7.2</td>
<td>1.9</td>
<td>0.4</td>
<td>7.7</td>
<td>4.0</td>
<td>13</td>
<td>42</td>
</tr>
<tr>
<td>8-Jun</td>
<td>Wholesale trade</td>
<td>4.6</td>
<td>3.4</td>
<td>4.0</td>
<td>6.4</td>
<td>7.2</td>
<td>1.9</td>
<td>0.4</td>
<td>7.7</td>
<td>4.1</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>12-Jun</td>
<td>CPI, Monthly Treasury Statement</td>
<td>4.6</td>
<td>3.4</td>
<td>4.0</td>
<td>6.4</td>
<td>7.2</td>
<td>1.6</td>
<td>0.8</td>
<td>7.7</td>
<td>4.0</td>
<td>13</td>
<td>45</td>
</tr>
<tr>
<td>13-Jun</td>
<td>Producer Price Index</td>
<td>4.6</td>
<td>3.4</td>
<td>4.0</td>
<td>6.4</td>
<td>7.3</td>
<td>1.4</td>
<td>0.9</td>
<td>7.7</td>
<td>4.0</td>
<td>13</td>
<td>44</td>
</tr>
<tr>
<td>14-Jun</td>
<td>Retail trade, Import/Export prices</td>
<td>4.8</td>
<td>3.6</td>
<td>3.9</td>
<td>6.4</td>
<td>7.3</td>
<td>1.4</td>
<td>0.9</td>
<td>7.8</td>
<td>3.8</td>
<td>15</td>
<td>44</td>
</tr>
<tr>
<td>15-Jun</td>
<td>Industrial production</td>
<td>4.6</td>
<td>3.7</td>
<td>2.9</td>
<td>6.4</td>
<td>7.4</td>
<td>0.3</td>
<td>0.8</td>
<td>7.1</td>
<td>3.1</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>19-Jun</td>
<td>Housing starts</td>
<td>4.7</td>
<td>3.7</td>
<td>2.9</td>
<td>6.4</td>
<td>7.5</td>
<td>2.9</td>
<td>0.8</td>
<td>7.1</td>
<td>3.1</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>20-Jun</td>
<td>Existing-home sales</td>
<td>4.7</td>
<td>3.7</td>
<td>2.9</td>
<td>6.4</td>
<td>7.5</td>
<td>2.2</td>
<td>0.8</td>
<td>7.1</td>
<td>3.1</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>25-Jun</td>
<td>New-home sales/prices/costs</td>
<td>4.6</td>
<td>3.7</td>
<td>2.9</td>
<td>6.4</td>
<td>7.6</td>
<td>0.6</td>
<td>0.9</td>
<td>7.1</td>
<td>3.1</td>
<td>17</td>
<td>37</td>
</tr>
<tr>
<td>27-Jun</td>
<td>Advance durable manufacturing, Advance Economic Indicators</td>
<td>4.5</td>
<td>3.7</td>
<td>2.9</td>
<td>6.4</td>
<td>7.6</td>
<td>0.6</td>
<td>0.9</td>
<td>7.9</td>
<td>1.9</td>
<td>29</td>
<td>20</td>
</tr>
<tr>
<td>29-Jun</td>
<td>GDP (6/28), Personal income and outlays, NPA underlying detail tables</td>
<td>3.8</td>
<td>2.7</td>
<td>4.0</td>
<td>6.9</td>
<td>7.6</td>
<td>0.6</td>
<td>0.8</td>
<td>8.0</td>
<td>2.2</td>
<td>28</td>
<td>19</td>
</tr>
<tr>
<td>2-Jul</td>
<td>ISM Manuf., Construction spending</td>
<td>4.1</td>
<td>2.9</td>
<td>4.8</td>
<td>6.9</td>
<td>5.3</td>
<td>0.9</td>
<td>1.6</td>
<td>8.4</td>
<td>2.6</td>
<td>27</td>
<td>19</td>
</tr>
</tbody>
</table>

**Maximum forecast of real GDP growth**
- Employment sit., ISM Manufacturing, 1-Jun
  - Construction spending: 4.8, 3.5, 6.1, 6.4, 7.2, 1.8, 0.4, 7.5, 3.2, 18, 43

**Minimum forecast of real GDP growth**
- GDP (6/28), Personal income and 29-Jun outlays, NPA underlying detail tables
  - 3.8, 2.7, 4.0, 6.9, 7.6, 0.6, 0.8, 8.0, 2.2, 28, 19

Note: CPI is “change in private inventories.” Changes in net exports and CPI are both in billions of 2009 dollars (SAAR). All other numbers are quarterly percent changes (SAAR). Table does not necessarily include all estimates for the quarter; see tab “TrackingHistory” in the online excel file for the entire history.
Big Data for Central Banking
Some use cases by MoF.

- Railways ticketing data (Internal Migration)
- Satellite data (Urbanization)
- GSTN data (Interstate commerce)
- Is this Big Data?
For central banks, the flexibility and real-time availability of big data open up the possibility of extracting more timely economic signals, applying new statistical methodologies, enhancing economic forecasts and financial stability assessments, and obtaining rapid feedback on policy impacts.

Economic science, political science, social science, computer science?
It is critical for a full-service Central Bank, such as the RBI, with diverse responsibilities – inflation management, currency management, debt management, reserves management, banking regulation and supervision, financial inclusion, financial market intelligence and analysis, and overall financial stability – to employ relevant data and apply the right filters for improving its forecasting, nowcasting, surveillance and early-warning detection abilities that all aid policy formulation. In the backdrop of ongoing explosion in information gathering, computing capability and analytical toolkits, policy making benefits not only from data collected through regulatory returns and surveys but also from large volumes of structured and unstructured real-time information sourced from consumer interactions in the digital world. Accordingly, it has been decided to gainfully harness the power of Big Data analytics by setting up a Data Sciences Lab within the RBI that will comprise experts and budding analysts, internal as well as lateral, who are trained inter alia in Computer Science, Data Analytics, Statistics, Economics, Econometrics and/or Finance. It is envisaged that the unit will become operational by December 2018.

Apr 2018 RBI press release
RBI’s Vision of CIMS

i) Tier-I: Data Collection Layer (DCL)
ii) Tier-II: Data Processing Layer (DPL)
iii) Tier-III: Final Data Repository Layer (FDRL)
iv) Tier-IV: Dissemination and Analytics Layer (DAL)
CIMS Major Components

From draft EoI

• System-to-System *automated data collection* of Macro and Micro data from source systems of banks and other entities to CIMS.

• Staging Area Data Portal (SADP) system for accessing/authorization/revision of data by banks and entities.

• Designed Data: Relational Database Management Systems (RDBMS) based repository with fact and dimension tables.

• Organic Data: Data Lake - which is composed of commodity hardware (one time write and multiple reads) as micro level big-data are stored in both structured and unstructured forms.

• Centralised Analytics Server will facilitate *user friendly* analytics and *multi-dimensional view* of data, its trends and relationships/correlations and the like.
Big Data for Central Banking


https://en.wikipedia.org/wiki/Big_data
The size of the World Wide Web (The Internet)

The Indexed Web contains **at least 4.49 billion pages** (Tuesday, 03 July, 2018).
The Dutch Indexed Web contains **at least 154.3 million pages** (Tuesday, 03 July, 2018).

The Indexed Web | The Dutch Indexed Web

---

Big Data for Central Banking
<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Source</th>
<th>Searches</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pak vs Aus</td>
<td>Cricbuzz</td>
<td>1M+</td>
</tr>
<tr>
<td>2</td>
<td>Aus vs Zim</td>
<td>Firstpost</td>
<td>500K+</td>
</tr>
<tr>
<td>3</td>
<td>Andheri Bridge Collapse</td>
<td>Mumbai rains Live: Local train services to be restored completely by tonight...</td>
<td>200K+</td>
</tr>
<tr>
<td>4</td>
<td>DHE Odisha</td>
<td>Hindustan Times</td>
<td>50K+</td>
</tr>
<tr>
<td>5</td>
<td>Mexico</td>
<td>Los Angeles Times</td>
<td>20K+</td>
</tr>
<tr>
<td>6</td>
<td>Sara Ali Khan</td>
<td>NDTV</td>
<td>20K+</td>
</tr>
<tr>
<td>7</td>
<td>CTET 2018 online application</td>
<td>The Indian Express</td>
<td>20K+</td>
</tr>
</tbody>
</table>
Big Data for Central Banking
Big Data for Central Banking
Homo *Erectus*, Homo *Sapiens*, Homo *Deus* [Yuval Noah Harari]

Change due to *Adaptation ≠ Improvement!*

**Technology (Wikipedia Definition)**

Technology is the usage and knowledge of tools, techniques, crafts, systems or methods of organization in order to *solve a problem* or *serve some purpose*.

Zero, Wheel, Printing Press, Radio, Lasers, ...

*Any sufficiently advanced technology is indistinguishable from magic.* [Arthur C. Clarke]

- **Why Information Technology is different?**
  Transistor, VLSI, Microprocessor, ...

- **Danger:** Computers are coming! Taking away our jobs!
  Construction, Farming, Banking, Surgery, Composing music, Teaching!
  Be very scared!
## Web 1.0, Web 2.0, Web 3.0

### Web 1.0 [1990-2005] (*Right to Information*)
- Internet: Info anytime, anywhere, any form
- Like *drinking water from a fire hose*
- Search Engines to the rescue

### Web 2.0 [2005-2015] (*Right to Assembly*)
- Social Networking (Twitter, Facebook, Kolaveri, Flash crowds)
- Producers, not only consumers (Wikipedia, blogs, ...)
- *Proliferated unreliable, contradictory information?*
- *Facilitated malicious uses including loss of privacy, security.*

### Web 3.0 [current] (*AI & ML meet Semantic Web*)
- Intelligent Agents that “understand”
- What do you want when you get up and put on computer?
- *I have a dream!* (MLK)
The **application layer** tapping the **hardware** (Web 1.0) and **OS** (Web 2.0)?

**Ramana Maharishi**

- **Lived** 30/12/1879 to 14/4/1950
- **Place:** Tiruvannamalai, Tamil Nadu

**contemporaries**

- Jiddu Krishnamurti
- Kanchi Chandrasekara Saraswathi

**author-of**

- Naan Yaar?
- Vichara Mani Mala
- Reality in Forty Verses

**Ramana Maharishi**

**Place:** Tiruvanamalai, Tamil Nadu

**Lived** 30/12/1879 to 14/4/1950

**contemporaries**

- Jiddu Krishnamurti
- Kanchi Chandrasekara Saraswathi

**author-of**

- Naan Yaar?
- Vichara Mani Mala
- Reality in Forty Verses

**Combined, categorized information inferred from various sites, languages.** [www.dbpedia.org](http://www.dbpedia.org) comes close today!
3rd platform: SMAC + IoT

- Main Frame (1960s ...)
- Client Server (1990s ...)
- Today (Handheld, Pervasive Computing)
3rd platform: SMAC + IoT

- What’s App (how many engineers?)
- Facebook, Twitter, GooglePlus …
- Web 2.0 (Right to Assembly)
- Crowdsourcing (Wikipedia)
- Crowdfunding (no banks!)

Social

3rd Platform

G. Sivakumar
Computer Science and Engineering

Big Data for Central Banking
3rd platform: SMAC + IoT

- Phone (Smart, Not-so-smart!)
- Wearables! (Google glass, Haptic)
- Internet of “Me” (highly personalized) Business (no generic products!)
- **BYOx:** Device security, App/content management nightmare.
- **Data Loss Prevention** (Fortress Approach - Firewall, IDS/IPS - won’t work!)

G. Sivakumar
Computer Science and Engineering

**Big Data for Central Banking**
3rd platform: SMAC + IoT

- Big Data Analytics
- Volume, Variety, Velocity, Veracity
- ACID properties Database not needed
- Hadoop, Map Reduce, NoSql
- Knowledge is Power!
- Collect, Analyse, Infer, Predict

G. Sivakumar
Computer Science and Engineering
Big Data for Central Banking
3rd platform: SMAC + IoT

- Moore’s law
- What could fit in a building .. room ... pocket ... blood cell!
- **Containers** Analogy from Shipping
  - VMs separate OS from bare metal (at great cost-
    Hypervisor, OS image)
  - **Docker**- separates apps from OS/infra using containers.
  - Like **IaaS, PaaS, SaaS**
  Have you heard of **CaaS**?

Big Data for Central Banking
3rd platform: SMAC + IoT

- **Sensors** (Location, Temperature, Motion, Sound, Vibration, Pressure, Current, ....)
- **Device Eco System** (Smart Phones, Communicate with so many servers!)
- **Ambient Services** (Maps, Messaging, Traffic modelling and prediction, ...)
- **Business Use Cases** (Ola Cabs, Home Depot, Philips Healthcare, ...)
- Impact on wireless bandwidth, storage, analytics (velocity of BIG data, not size)
Can AI of computers match NS of humans?

Old Joke: *Out of sight, out of mind*

Consider chess, once the *holy grail* of AI. Praggnaananda (prodigy!)

Does not play the human way at all! Mostly parallelized search in hardware (200 million positions/second!)

December 2017: *AlphaGo Zero* used *reinforcement learning* to teach itself chess in 4 hours! Beat world’s best program *Stockfish* comprehensively!
IIT Bombay motto: ज्ञानम् परमम् ध्येयम्
न चोरहायं न च राजहायं न आतुभाज्यम् न च भारकारी
व्यये कृते वर्धन एव नित्यं विद्यायनं सर्वधनप्रधानं

*It cannot be stolen by thieves, cannot be taken away by the king, cannot be divided among brothers and does not cause a load. If spent, it always multiplies. The wealth of knowledge is the greatest among all wealths.*

- How is *learning* affected by the *information deluge*?
- Data is not Information/Knowledge/Wisdom.
- Analogy of Monkeys randomly typing
- How to find Shakespeare’s verse?
- Data Mining (My-ning? ;-)
- Has Internet damaged the human brain?

**Big Data for Central Banking**

G. Sivakumar
Computer Science and Engineering

शिवकुमार ग. सिवकुमार
मध्यवर्ती संविदा संस्थान मुंबई
https://en.wikipedia.org/wiki/Data_science

- Turing award winner Jim Gray terms it the 4th paradigm after Empirical, Theoretical and Computational
- How does Google translate documents?
- Combining data sources to produce new information not contained in any single one!
- How does Alpha-Zero play chess?
- Deep Learning! (Speech Recognition)
Are doctors practicing medical science?

https://www.nature.com/articles/srep26094

The machine was given no information about how the human body works or how diseases affect us. It found correlations that let it predict the onset of some diseases more accurately than ever, and some diseases, such as schizophrenia, for the first time at all. It does this by creating a vast network of weighted connections that is just too complex for us to understand.
4-Vs of Big Data (IBM’s Version)

The FOUR V’s of Big Data

Volume: SCALE OF DATA

- It’s estimated that 2.5 quintillion bytes of data are created each day.
- 2020: 40 zettabytes (40 trillion gigabytes) of data will be created by 2020, an increase of 300 times from 2005.
- 6 billion people have cell phones.
- World population: 7 billion.

Variety: DIFFERENT FORMS OF DATA

- As of 2011, the global size of data in healthcare was estimated to be 150 exabytes (150 trillion gigabytes).
- 420 million wearable, wireless health monitors.
- 4 billion+ hours of video are watched on YouTube each month.
- 30 billion pieces of content are shared on Facebook every month.
- 400 million tweets are sent per day by about 200 million monthly active users.

Velocity: ANALYSIS OF STREAMING DATA

- Modern cars have close to 100 sensors that monitor items such as fuel level and tire pressure.
- The New York Stock Exchange captures 1 TB of trade information during each trading session.
- By 2015, it is projected there will be 18.9 billion network connections — almost 2.5 connections per person on earth.

Veracity: UNCERTAINTY OF DATA

- 27% of respondents are unsure of how much of their data was inaccurate.
- 1 in 3 business leaders don’t trust the information they use to make decisions.
- Poor data quality costs the US economy around $3.1 trillion a year.

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAG, IBM, MIPTEC, QAS.
So, What’s Big Data?

Big Data for Central Banking
Big Data Technology Stack

- Data Ingestion
  - Extract Transform Load (ETL)
  - continuous or asynchronous, real-time or batched
  - Flume/Kafka

- Data Storage
  - Data Center, Cluster Nodes (Commodity Hardware)
  - HDFS, S3, HBase, Cassandra, MongoDB (document)

- Data Processing/Analytics
  - MapReduce, Spark (in-memory), Storm,
  - R, Python, Java, Scala, ...
  - Mahout (machine learning)
  - Lambda Architecture

- Data Visualization
  - Reporting, Alert, Recommendation, Dashboard
  - Tableau, Kibana

To make all this work, we need Security, Orchestration
Layered Technology Stack

Big Data Applications
Reporting and Visualization
Analytics (Traditional and Advanced)
Analytical Data Warehouses and Data Marts
“Organizing” Databases and Tools
Operational Databases (Structured, Unstructured, Semi-Structured)
Security Infrastructure
Redundant Physical Infrastructure

Interfaces and feeds from/to the Internet
Interfaces and feeds from/to internal applications
Opportunity knocks, but once.
There is a tide in the affairs of men, Which taken at the flood, leads on to fortune. Omitted, all the voyage of their life is bound in shallows and in miseries. *Shakespeare*

Forewarned is Forearmed.
The effect of disasters should be thought of beforehand. It is not appropriate to start digging a well when the house is ablaze with fire.
Past (What happened? Reactive)
- Designed Batch/Static Data Reports, Standards, Data Harmonization.

Present (What is happening?)
- Organic Unstructured Streaming/Real-time Data Statistical Analysis, Anamolies, Alerts

Future (What will happen? Pro-active)
- Predictive Forecast, Optimize

Analytics can convert *data to knowledge to wisdom*.
What’s special in Big Data Analytics?

What more/different do we need from

- **Databases (RDBMS)**
  Used for Transaction processing, consolidation, reports, business performance

- **Data Warehouse**
  Used to Co-relate various data, strategies for pricing, supply chains, cross-sale, ....

Can these not handle Web log for profiling and recommendations, Sentiment analysis for product evaluation, positioning, marketing?

No. Biggest problem is the **ACID** properties requirement which makes *scale-out* and *fault-tolerance* near impossible.

Some Weaknesses:

- Disk oriented storage and indexing structures
- Multithreading to hide latency
- Locking-based concurrency control mechanisms
- Log-based recovery
Big Data Analytics Requirements

Think: Facebook likes, Shopping cart recommendations, ...

- Ingest data at very high speeds and rates
- Scale easily to meet growth and demand peaks
- Support integrated fault tolerance
- Support a wide range of real-time (or “near-real-time”) analytics
- Integrate easily with high volume analytic datastores
High Speed Data Ingestion

- Support millions of write operations per second at scale
- Read and write latencies below 50 milliseconds
- Do not need ACID-level consistency guarantees (Eventual is fine!)
- Support one or more well-known application interfaces
  - SQL
  - Key/Value
  - Documents (NLP)
Scaling Requirements

- Scale-out on commodity hardware
- Built-in database partitioning
  Manual sharding and/or add-on solutions not practical. Database must automatically implement defined partitioning strategy.
- Application should see a single database instance
- Database should encourage scalability best practices
  For example, replication of reference data minimizes need for multi-partition operations.
Data Governance

Most critical, but most neglected (like in this talk!)

FIGURE 1: The Data Life Cycle and Surrounding Data Ecosystem

from the National Science Foundation CISE
AC Data Science Report, October 2016
Know Thyself (more than the Algorithms know you)

_The algorithms are watching you right now. They are watching where you go, what you buy, who you meet. Soon they will monitor all your steps, all your breaths, all your heartbeats. They are relying on Big Data and machine learning to get to know you better and better. And once these algorithms know you better than you know yourself, they could control and manipulate you, and you won’t be able to do much about it. You will live in the matrix ... Yuval Noah Harari_
RoI:

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

What next?

RoI:

आचार्यांत पादमादते पादं शिष्यं स्वमेधया ।
सवल्लचारिज्यं पादं पादं कालक्रमेण च ॥

one fourth from the teacher,
one fourth from own intelligence,
one fourth from classmates,
and one fourth only with time.
Some useful resources.

- [https://www.bis.org/ifc/publ/ifcb44_overview_rh.pdf](https://www.bis.org/ifc/publ/ifcb44_overview_rh.pdf) (Summary of the proceedings of a recent workshop on Big Data for Central Banks)
- [https://www.newyorkfed.org/research/policy/nowcast](https://www.newyorkfed.org/research/policy/nowcast)
- [https://www.frbatlanta.org/cqer/research/gdpnow.aspx](https://www.frbatlanta.org/cqer/research/gdpnow.aspx)
- Relevant *Wikipedia* pages