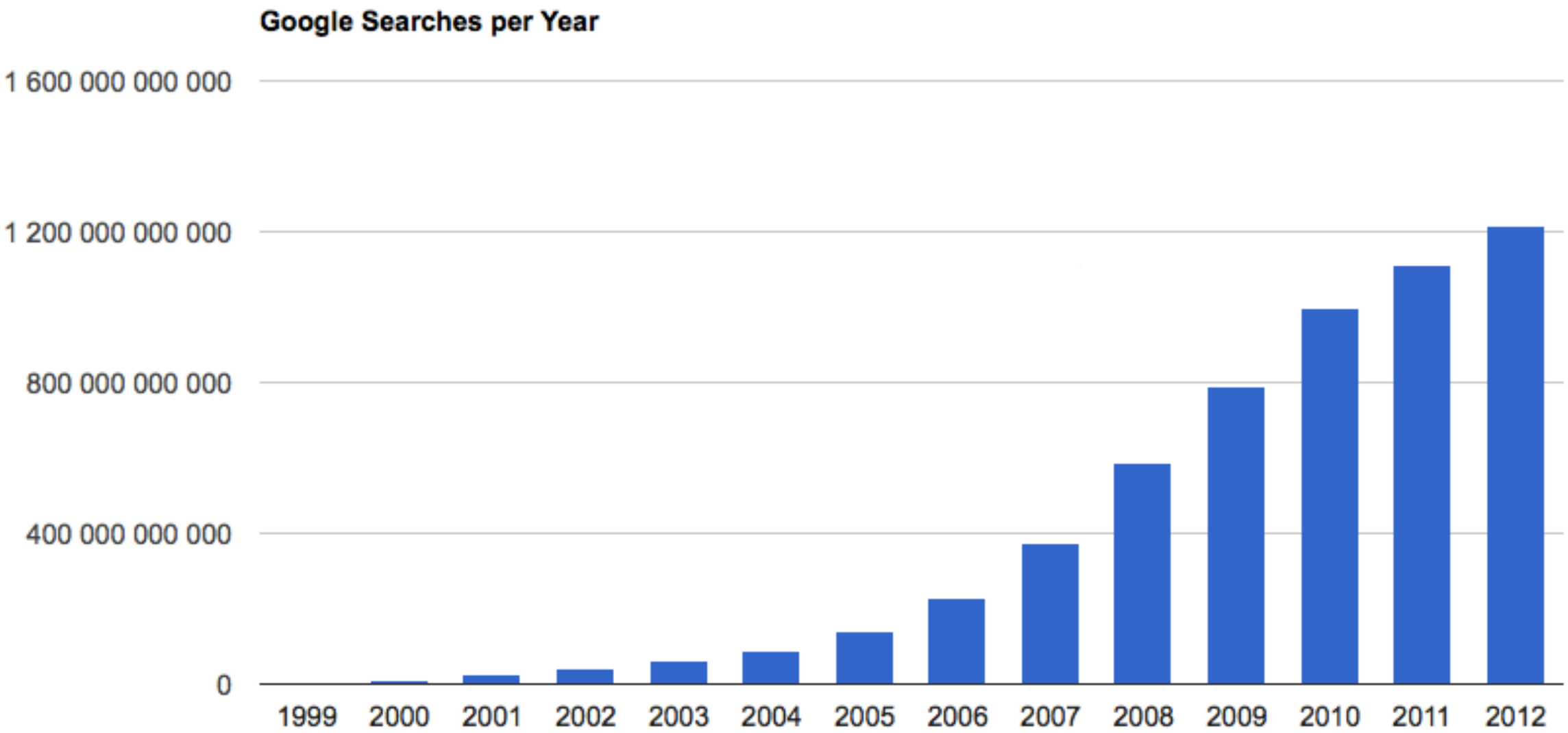
Big data in official statistics

Dominik Rozkrut
Główny Urząd Statystyczny

Big data

Google searches

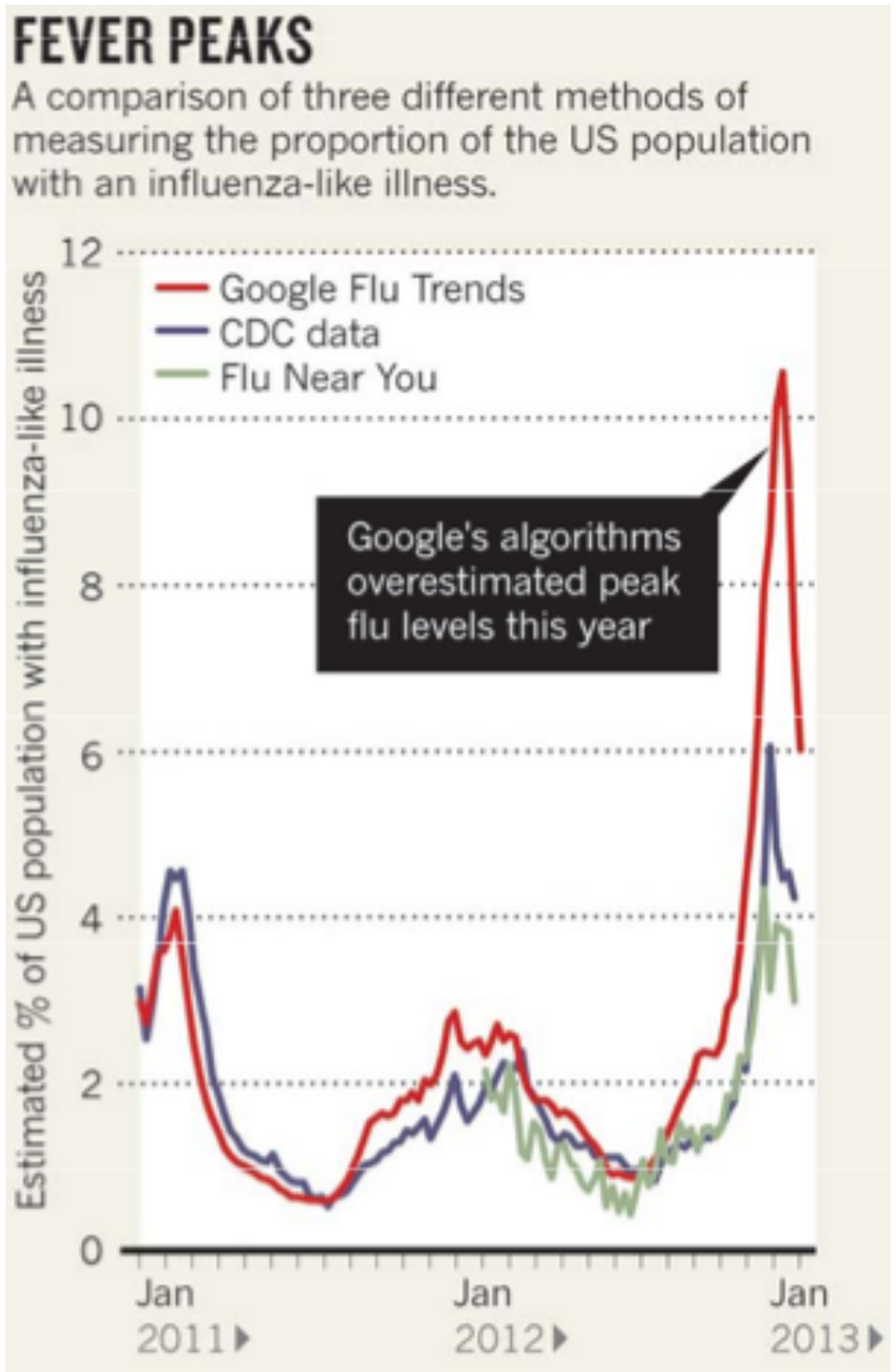


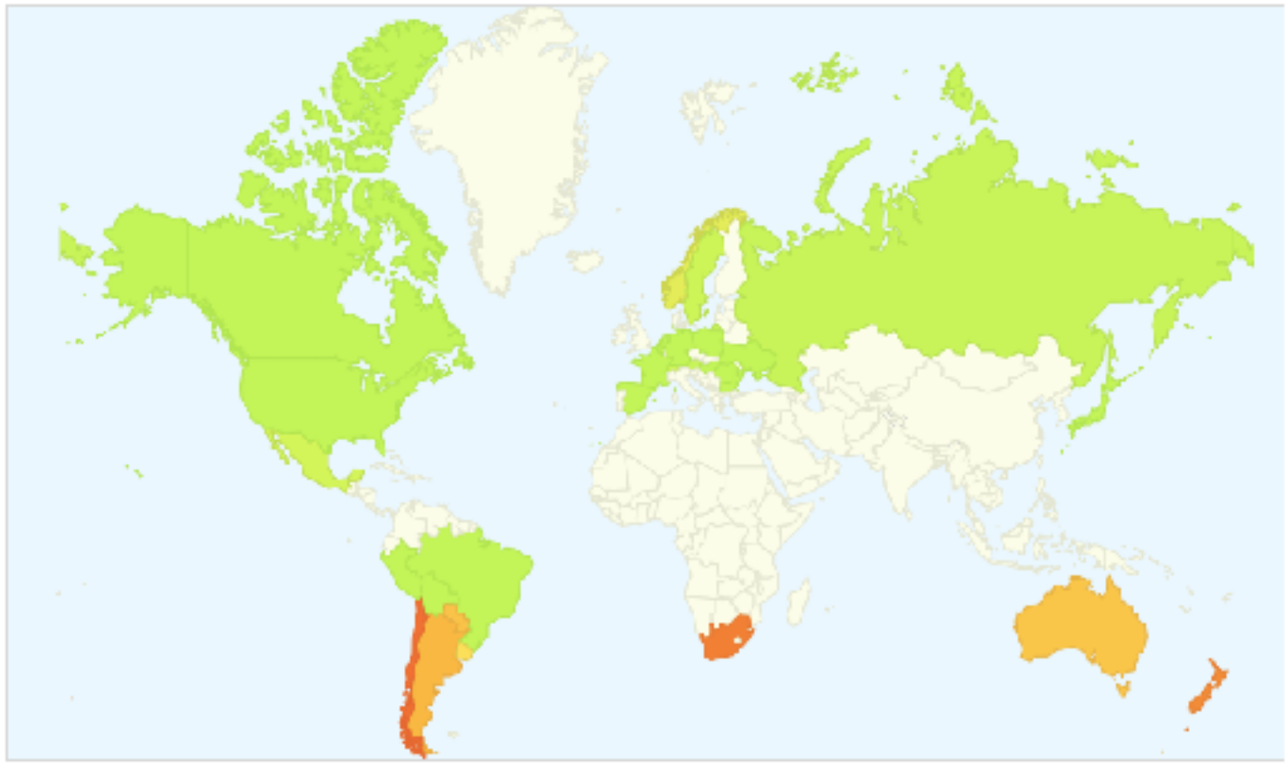
Most popular web sites

	SE Search engine	SN Social network	MC Media-Content	PT Portal	Ref. Ref.-Encyclopedia	Ns News	e$ eComm/banking	Blog Blogging

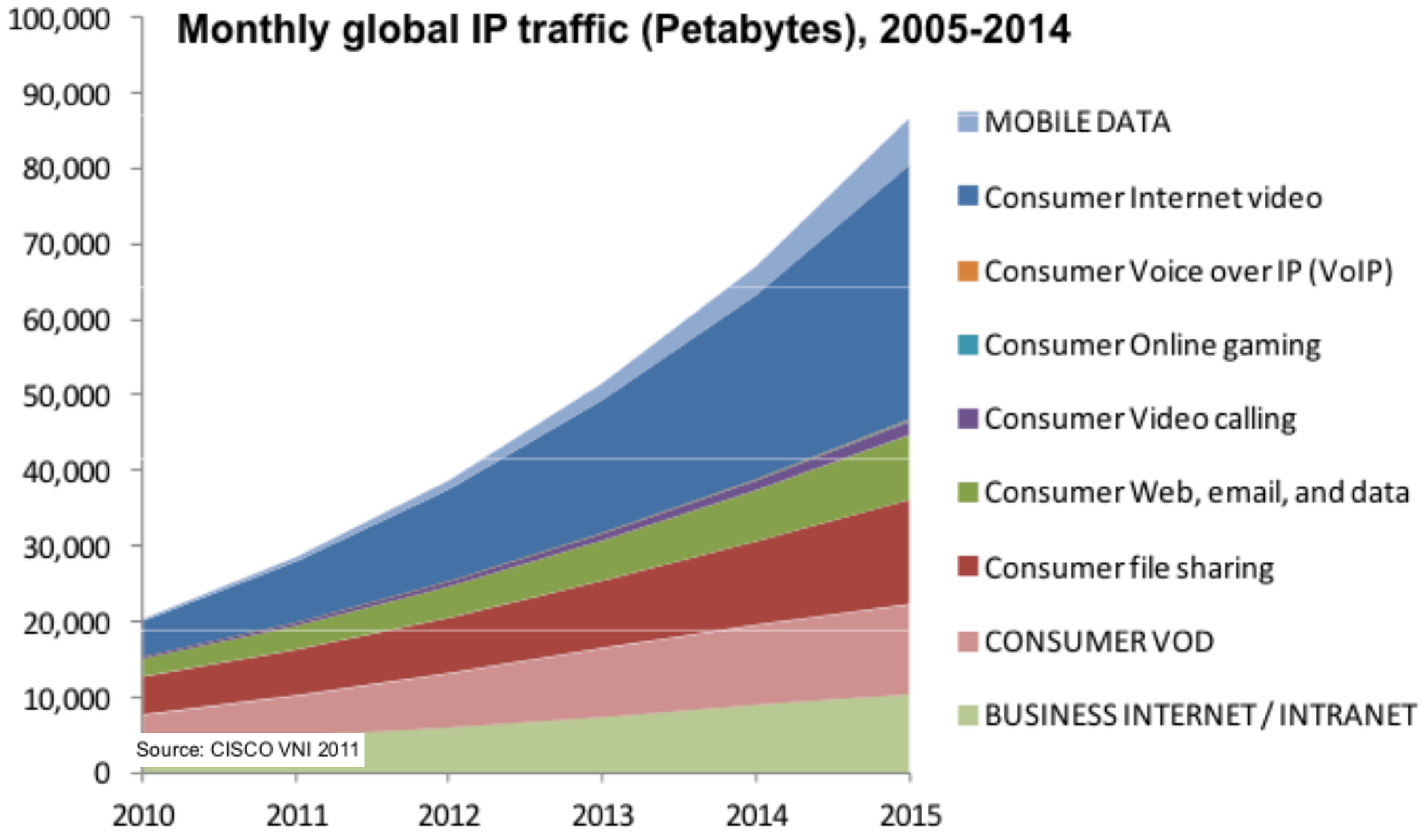
Rank	AUS	AUT	BEL	CAN	CHL	CZE	DNK	EST	FIN	FRA	DEU	GRC	HUN	ISL	IRL	ISR	ITA	JPN	KOR	LUX	MEX	NLD	NZL	NOR	POL	PRT	SVK	SVN	ESP	SWE	CHE	TUR	GBR	USA
1	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SN	SE	SE	SE	PT	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE	SE
2	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SE	SN	SN	SN	SE	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN	SN
3	MC	MC	MC	MC	MC	PT	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	PT	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC	MC
4	PT	MC	Ref.	PT	PT	MC	SN	PT	Ref.	PT	e$.	e$.	MC	MC	SN	SN	PT	e$.	MC	Ref.	PT	Ref.	e$.	MC	PT	PT	PT	MC	MC	Ref.	Ref.	Ns	MC	PT
5	e$.	Ref.	PT	Ref.	PT	MC	Ref.	MC	Ns	Ref.	e$.	PT	e$.	MC	Ref.	PT	Ref.	PT	PT	PT	PT	PT	PT	e$.	e$.	e$.	Ns	MC	PT	Ns	PT	Ns	e$.	e$.
6	Ref.	e$.	SN	SN	MC	MC	Ns	Ref.	Ns	e$.	Ref.	Ref.	Ns	Ns	PT	Ref.	PT	SN	PT	SE	MC	SN	Ref.	Ref.	PT	Ref.	MC	PT	SN	e$.	SN	MC	PT	SN
7	SN	Ns	PT	PT	Ref.	Ref.	MC	PT	MC	PT	Ns	PT	Ref.	MC	MC	PT	Ns	e$.	PT	SN	Ref.	PT	Ns	Ns	Blog	PT	PT	Ref.	Ref.	PT	MC	PT	Ref.	Ref.
8	PT	e$.	MC	MC	Blog	MC	PT	SN	MC	e$.	PT	MC	PT	PT	PT	MC	e$.	Ref.	PT	e$.	PT	PT	MC	MC	MC	MC	Ns	e$.	PT	PT	Ns	e$.	e$.	e$.
9	Ns	SE	MC	e$.	SN	MC	PT	e$	e$	PT	Ns	MC	MC	MC	e$	PT	Blog	PT	PT	PT	e$.	e$.	SN	PT	Ref.	SN	MC	MC	Blog	SN	e$.	MC	SN	MC
10	e$.	PT	SE	e$.	Ns	e$.	e$.	PT	PT	SN	PT	Blog	PT	Ref.	e$.	e$.	e$.	PT	PT	MC	SN	MC	PT	PT	SE	PT	Ref.	e$.	Ns	e$	PT	Blog	PT	SE

Google Flu Trends

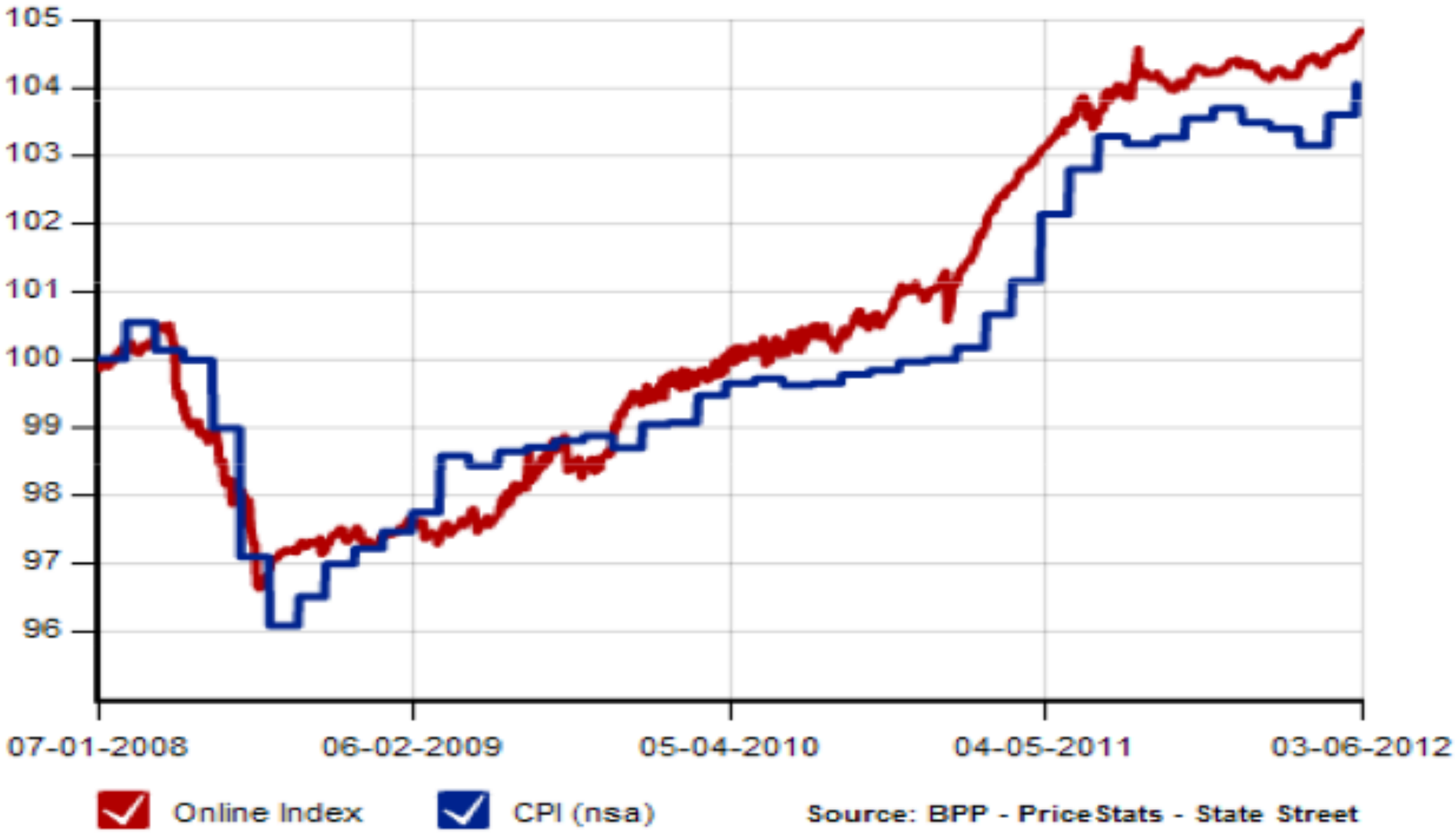




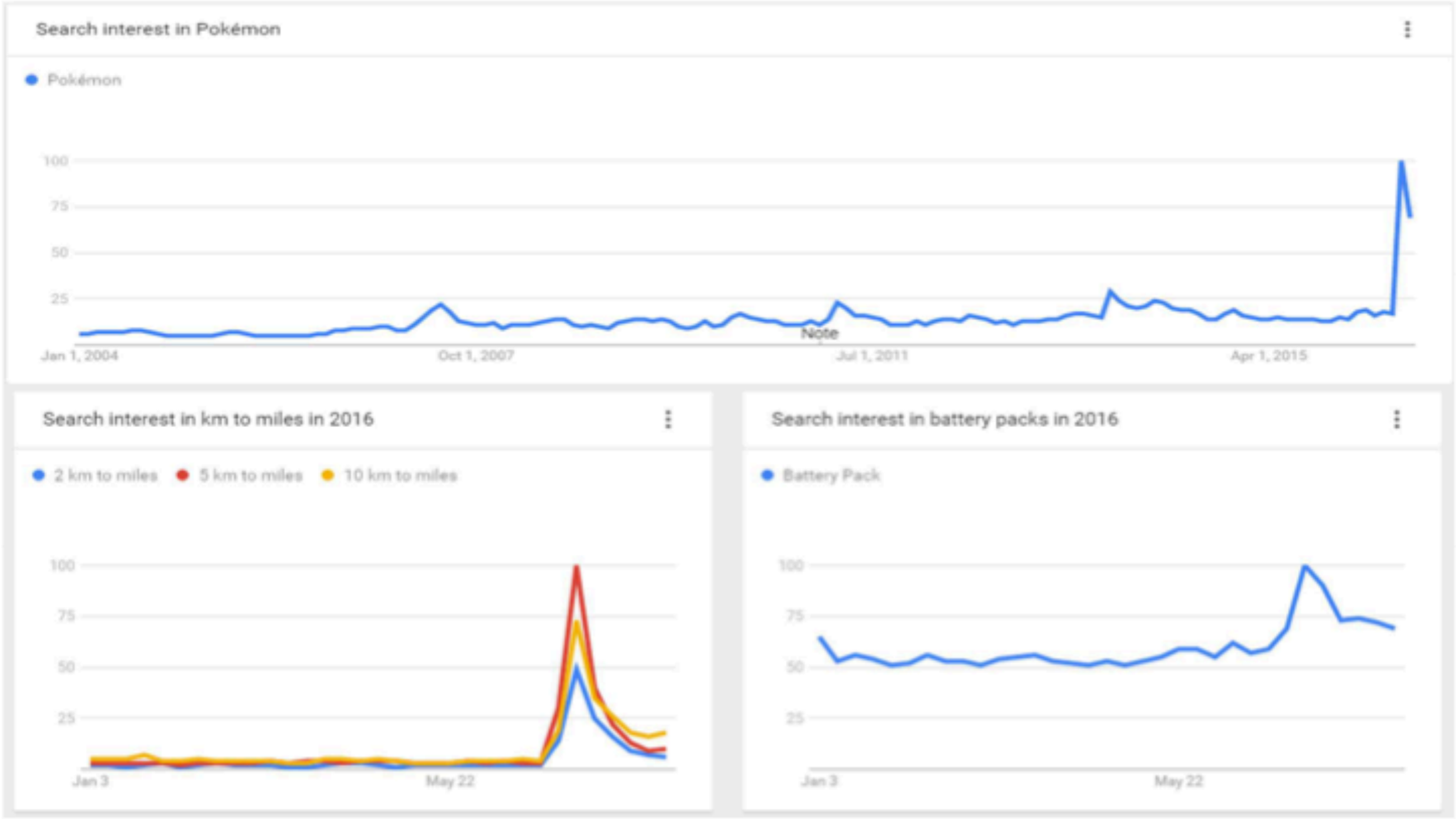
Global IP traffic



MIT Billion Price Project



Pokemon


Search interest in Pokémon
Pokémon
100
75
50
25
Note
Jan 1, 2004
Oct 1, 2007
Jul 1, 2011
Apr 1, 2015
Search interest in km to miles in 2016
2 km to miles
5 km to miles
10 km to miles
100
75
50
25
Jan 3
May 22
Search interest in battery packs in 2016
Battery Pack
100
75
50
25
Jan 3
May 22

Lots of data is being collected and warehoused

- web data,
- e-commerce
- purchases at department/grocery stores
- bank/credit card transactions
- social networks
- …

How much?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN's Large Hydron Collider (LHC) generates 15 PB a year

Problem?

- data is growing at a tremendous rate

- however the increase in data is a minor problem, the increase in percentage of unstructured data in the overall data volume is what is concerning all

- a common definition across sectors for big data is a strategy or technology deployment that deals with data problems that are too large, too fast, or too complex for conventional database or processing technology

- the manipulation, management, and interpretation of these data sets —complex, high volume, or varied — therefore constitutes a big data challenge, according to this logic

Introduction to big data

- experiments, observations, and numerical simulations in many areas of science and business are currently generating terabytes of data, and in some cases are on the verge of generating petabytes

- traditional methods of analysis have been based largely on the assumption that analysts can work with data within the confines of their own computing environment, but the growth of “big data” is changing that paradigm, especially in cases in which massive amounts of data are distributed across locations

- Google, Yahoo!, Microsoft, and other Internet-based companies have data that is measured in exabytes

Introduction to big data

- social media (e.g., Facebook, YouTube, Twitter) have exploded beyond anyone's wildest imagination, and today some of these companies have hundreds of millions of users

- innovations in the fields of machine learning, data mining, statistics, and the theory of algorithms have yielded

- data analysis methods that can be applied to ever-larger data sets

- data mining of these massive data sets is transforming the way we think about crisis response, marketing, entertainment, cybersecurity, national intelligence, information storage and retrieval

Big Data is not new

- 1997 - a paper that discussed the difficulty of visualizing Big Data

- 1999 - a paper that discussed the problems of gaining insight from the numbers in Big Data

- potential sources of discovery and knowledge, requiring sophisticated analysis techniques that go far beyond classical indexing and keyword counting, aiming to find relational and semantic interpretations of the phenomena underlying the data.

- big data is increasingly becoming a factor in production, market competitiveness and, therefore, growth

Definition

- the concept of big data is therefore often linked to the three Vs, which were identified by Gartner in 2001:
  - **Velocity**—the speed of data delivery and processing
  - **Volume**—the amount of data that must be managed or processed
  - **Variety**—the range of different data sets that must be dealt with, including both structured and unstructured formats
- another V has been added to the list by some academics:
  - **Veracity**, which relates to measurement or labeling of the integrity or quality (… variability, complexity…)

Data management

- the traditional RDBMS (Relational Database Modeling Systems) cannot handle big data and hence there is a shift towards specialized, non-relational databases like Hadoop

- SQL is being replaced by Map Reduce, which is the distributed querying and data processing engine used to extract data from big datasets hosted on clusters in any typical Hadoop implementation

Challenges

- in the context of massive data care must be taken with all such tools for two main reasons:

  - all statistical tools are based on assumptions about characteristics of the data set and the way it was sampled, and those assumptions may be violated in the process of assembling massive data sets

  - tools for assessing errors of procedures, and for diagnostics, are themselves computational procedures that may be computationally infeasible as data sets move into the massive scale

Challenges - quality of data

- the word “data” connotes fixed numbers inside hard grids of information, and as a result, it is easily mistaken for fact

- bad data causing big mistakes

- big data raises bigger issues - there are even more hazards, some human and some inherent in the technology

- big data isn’t objective - most data sets, particularly where people are concerned, need references to the context in which they were created

- big data is only as good as the people using it

Big data in official statistics

IT has always had an impact on Statistics

Types of data (e.g.)

- mobile phone data
- satellite imagery
- social media data
- web-scraping data
- sensors data
- …

Applications (e.g.)

- transportation statistics (traffic sensors)
- mobility statistics (mobile phones)
- business statistics (web scrapping)
- environmental statistics
  - AIS data
  - satellite imagery and geospatial data to monitor the environment at high temporal and spatial resolutions

Benefits

- faster, more timely statistics (timeliness & now-casting)
- new products and services
- cost reduction, affordability
- robustness and granularity
- democratization and creativity

Issues

- training, skills and capacity-building
- methodology
- classifications
- quality frameworks
- access to and ownership of proprietary data liability
- quick technology change

Technology

- Apache Hadoop
- Apache Spark
- Cassandra
- Redis, Riak
- MongoDB
- Apache Nutch
- HTTrack
- jsoup
- Apache SOLR
- …

Methodological challenges

- unknown bias
- potential instability
- quality

Skills

- methodologist on big data issues
- data scientist
- mathematical modeling specialist
- IT architecture specialist
- data visualization specialist
- cybersecurity specialist

What's needed now

- quality framework for big data
- skills and training for big data
- access to big data
- methodologies, estimation methods

Big data quality framework

- input quality (data access, sources)
- throughput quality (big data methodology and estimation methods, statistical production processes)
- output quality (application of big data to official statistics)

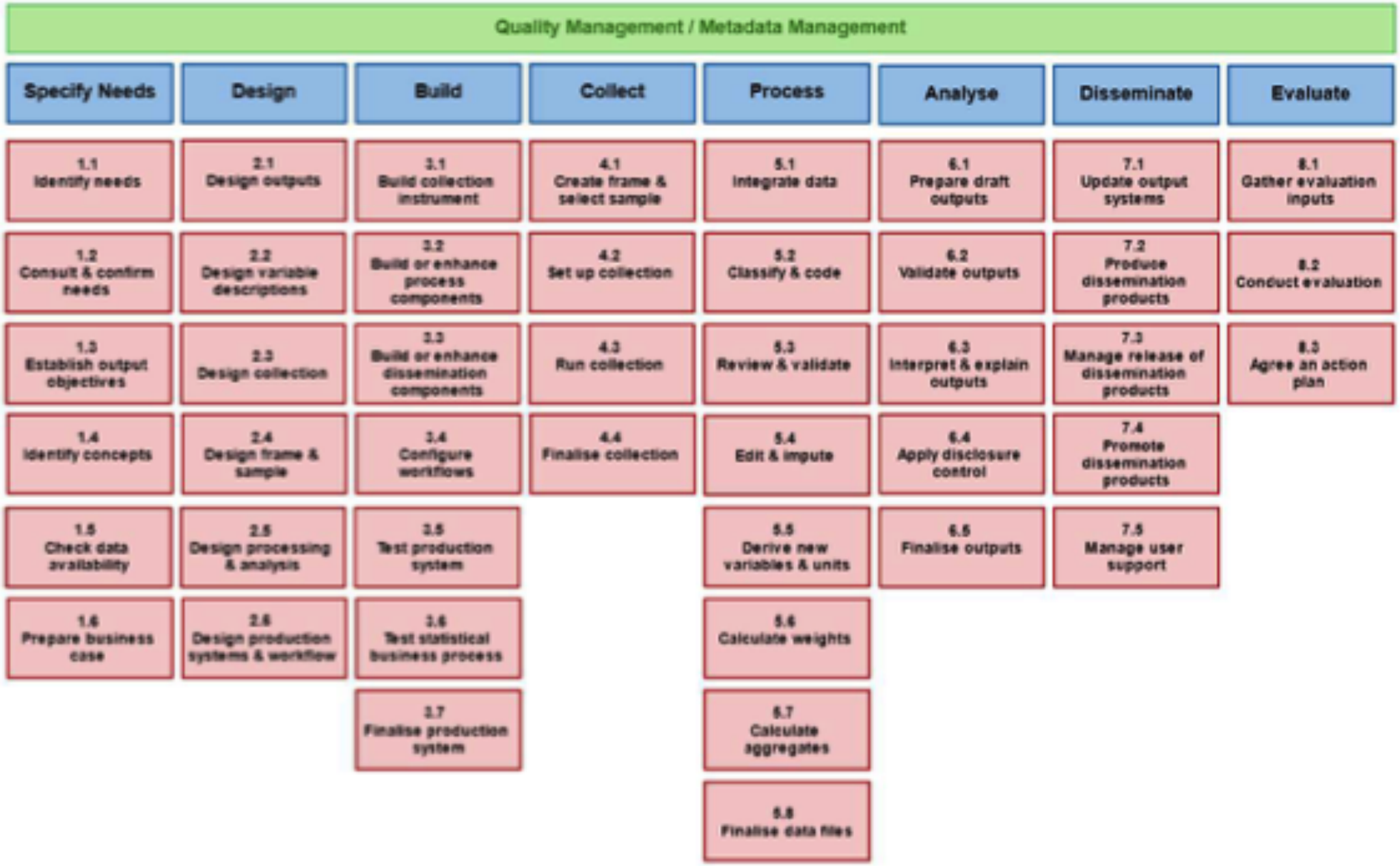
Big Data Project 2014

- 70 experts from national and international statistical organizations around the world

- overseen by the High-Level Group for the Modernization of Statistical Production and Services (HLG)

- the HLG was set up by the Bureau of the Conference of European Statisticians (CES) in 2010. CES reports to the United Nations Economic Commission for Europe (UNECE)

GSPBM


Quality Management / Metadata Management
Specify Needs
Design
Build
Collect
Process
Analyse
Disseminate
Evaluate
1.1 Identify needs
2.1 Design outputs
3.1 Build collection instrument
4.1 Create frame & select sample
5.1 Integrate data
6.1 Prepare draft outputs
7.1 Update output systems
8.1 Gather evaluation inputs
1.2 Consult & confirm needs
2.2 Design variable descriptions
3.2 Build or enhance process components
4.2 Set up collection
5.2 Classify & code
6.2 Validate outputs
7.2 Produce dissemination products
8.2 Conduct evaluation
1.3 Establish output objectives
2.3 Design collection
3.3 Build or enhance dissemination components
4.3 Run collection
5.3 Review & validate
6.3 Interpret & explain outputs
7.3 Manage release of dissemination products
8.3 Agree an action plan
1.4 Identify concepts
2.4 Design frame & sample
3.4 Configure workflows
4.4 Finalise collection
5.4 Edit & impute
6.4 Apply disclosure control
7.4 Promote dissemination products
1.5 Check data availability
2.5 Design processing & analysis
3.5 Test production system
5.5 Derive new variables & units
6.5 Finalise outputs
7.5 Manage user support
1.6 Prepare business case
2.6 Design production systems & workflow
3.6 Test statistical business process
5.6 Calculate weights
3.7 Finalise production system
5.7 Calculate aggregates
5.8 Finalise data files

Quality task team

- develop quality framework for big data
- build on existing frameworks developed for administrative data
- identify new challenges
- primary audience : National Statistical Institutes (NSIs) producing official statistics

Hyperdimensions

- source:
  - related to the type of data, the entity from which the data is obtained, and how it is administered and regulated.
- metadata:
  - description of concepts, file contents, and processes.
- data:
  - relates to quality of the data itself

Quality dimensions

- institutional environment
- privacy and security
- completeness, usability, time factors,
- accuracy (selectivity)
- coherence (linkability)
- validity
- accessibility, clarity, relevance

Institutional environment

- new data providers with evolving business models
- institutional and organizational factors
    - effectiveness and credibility of the organization
    - input phase: data provider
    - output phase: NSI
- transparency
    - assessment of the surrounding context, which may influence the validity, reliability or appropriateness of the product.
- longevity and stability of organization itself
    - potential to obtain similar products in the future, from the organization

Privacy and security

- legal limitations, confidentiality, security and privacy concerns
  - big data provider
  - organization that plans to use the data
  - Entity to which the information pertains (businesses, people, etc.,)
- data publicly available
- consent
  - active or passive
  - must meet NSI's guidelines, policies, and regulatory environment
- public trust
  - perceived lack of consent due to data acquisition may undermine public trust
  - mitigation actions may be necessary

Complexity

- context
    - affect receipt, storage, process and dissemination
- data structure
    - relational database
        - data integration of multiple sources (with or without linking variables among sources)
        - number and size of unstructured files that must be integrated
- data format
    - different data standards are used to store the data (e.g., spatial data mapped in various formats).
    - code lists used in the datasets are not harmonized (e.g., to code gender different labelling are used in different sources)

Usability

- ease to which the NSI will be able to work with and use the new data source
- specialized human resources
  - IT environment, big data analysis
- infrastructure for storing, processing, analysis, and dissemination
  - integration to existing infrastructure and standards
  - need to develop new infrastructure and standards
- capacity and cost
  - balance between expected gains and the cost

Time factors - timeliness and frequency

- Considered to be and the added value provided by Big Data
- Real-time analysis
- Challenges

- Time-related problems such as delays between the reference period (the point in time that the data refers to) and the time of collection;

- Reference period or the time of collection may not be known with certainty.

Accuracy

- the accuracy of statistical information is the degree to which the information correctly describes the phenomena it was designed to measure.

- recommend a total error approach

- challenges
  - metrics depends on the type of data
  - need to review or develop metrics for new data sources

Selectivity

- a key concern with many big data sources is the selectivity, (or conversely, the representativeness) of the dataset

- a dataset that is highly unrepresentative may nonetheless be useable for some purposes but inadequate for others

- related to this issue is the whether there exists the ability to calibrate the dataset or perform external validity checks using reference datasets

Coherence

- extent to which the dataset follows standard conventions, is internally consistent, is consistent over time, and is consistent with other data sources.

- linkability

    - Ease with which the data can be linked or merged with other relevant datasets.

- consistency

    - extent to which the dataset complies with standard definitions and is consistent over time.

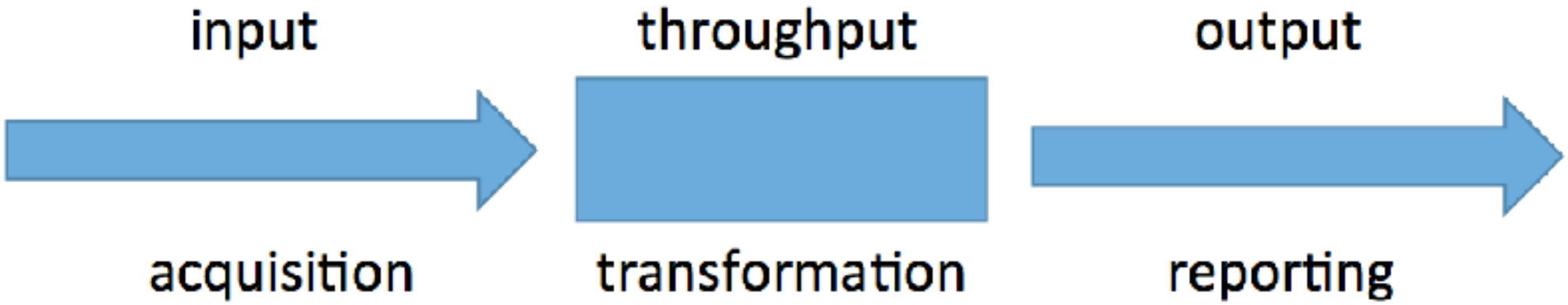
Validity

- the validity of a dataset is the extent to which it measures what the user is attempting to measure.

- the concept of validity is a long-standing one in methodology

Approach

- for each phase, define appropriate quality dimensions and quality indicators



Input

Hyperdi-mension	Quality Dimension	Factors to consider
Source	**Institutional Environment**	Sustainability of the entity-data provider Reliability status, transparency, interpretability
	Privacy and Security	Legislation, Data Keeper vs. Data provider Restrictions, Perception
Metadata	**Complexity**	Technical constraints, Sructured or Unstructured Readability, Presence of hierarchies and nesting
	Completeness	Metadata is available, interpretable and complete
	Usability	Resources required to import and analyse Risk analysis
	Time-related	Timeliness, Periodicity, Changes through time
	Linkability	Presence and quality of linking variables
	Coherence	Use of standards
	Validity	Transparency of methods and processes Soundness of methods and processes

Input

Hyperdimension	Quality Dimension	Factors to consider
Data	**Accuracy and selectivity**	Total error approach Reference datasets Selectivity
	Linkability	Quality of linking variables
	Coherence - consistency	Coherence between metadata description and observed data values
	Validity	Coherence between processes and methods and observed data values

Throughput

- system Independence: the result of processing the data should be independent of the hardware and software systems used to process it

- steady states: that the data be processed through a series of stable versions that can be referenced by future processes and by multiple parts of the organization

- application of quality gates: that the NSO employ quality gates as a quality control business process

Output

- Output quality framework should
  - Meet reporting criteria of the NSI
  - Provide required information to allow users to make informed decision regarding the use of the statistical output
- Follow transparency principle
- Follow the general approach with quality dimensions, indicators and factors to consider used for the input phase

Conclusions – Quality Task Team

- there is a need for quality assessment covering the entire business process

- a framework has been proposed for the Input and Output phases

- quality processing principles are proposed for the Throughput phase

Methodological issues cont.: sampling

- random sample gives an unbiased representation of the population
- big data can now be collected …
- … it requires sampling and testing methods that are beyond the state of the art
- … sampling implementation and estimation more difficult
- … some devices may be much more reliable than others

Good news

- big data tools are becoming widely available
- the cloud can address infrastructure needs

Big data needs official statistics as much as official statistics need big data

- official statistics is anchored in internationally agreed quality frameworks and methodologies

- official statistics is based on principles of professional independence and trust

- official statistics using traditional source data that allow methods for generating statistics from big data sources to be calibrated and validated

- statistical methodology can turn big data into small data, through sampling,

- transfer of data is not always necessary, as the method can be applied at the location of the data source

Communication

- building public trust in the use of big data from the private sector

- communicate the benefits and value of big data

Conclusions

- “The sexy job in the next 10 years will be statisticians.”
    - NYT, 5 August 2009
- “Data is the new oil.”
    - Andreas Weigend, Stanford
- “The future belongs to companies and people that turn data into products”
    - Mike Loukides, O’Reilly Media