Non-traditional data sources for social and economic monitoring

Josep Domenech
Universitat Politècnica de València

Introduction

- Digital Era: ICTs are in most of the everyday activities
  - Companies
    - Online marketing campaigns
    - Monitoring sales representatives with smartphones
    - Sensors in the workplace or in the machines
    - …
  - Individuals
    - Online purchases
    - Opinion sharing
    - Chat with friends
    - Get directions
    - …
- Measuring business and individual activities should not neglect the digital behavior

Introduction

- Digital footprint generated by people and business activities:
  - Potential for monitoring and revealing trends in economic, industrial and social behavior.
- Concept of Big Data early defined with the 3Vs model (Laney 2001):
  - Volume (size of data)
  - Velocity (speed of data transfers)
  - Variety (different types of data)
- Later extended with more Vs (Bello-Orgaz et al 2016):
  - Value (analytics)
  - Veracity (privacy and governance)

Introduction

- New data paradigm to transform the socio-economic policy and research as well as for business management and decision-making (Einav & Levin 2014, Varian 2014)
- To generate value, it is basic to:
  - Identify data sources
  - Know how to treat these data
  - Systematize the process

Background

- The technical roots of a Big Data architecture are in distributed computing paradigms such as grid computing (Berman et al. 2003)
- The Big Data analytics workflow includes (Assunçao et al 2015):
  - Data sources
  - Data management
  - Modelling
  - Result analysis and visualization
- Related to the cloud computing paradigm: fundamental for providing high data storage and computing power

Non-traditional sources of data

- These include the digital footprint left by individuals and companies
- Classification according to the purpose of the user when data are generated
  - Information search (Google Trends)
  - Information diffusion (websites, wiki pages)
  - Social interaction (SNS)
  - Transactions (retail scanners, card readers)
  - Non-deliberate (cookies, IP address, CDR, WIFI APs)

Non-traditional sources of data

- Websites
  - They are the public image of companies in the Internet
    - They inform about products, services, structure, intentions…
  - Classification of functions:
    - Informing
    - Conducting transactions
    - Facilitating opinion sharing
  - Research results related to the analysis of the *Informing* function:
    - Firms sales growth (Li et al. 2016)
    - Innovation and technology adoption (Shapira et al. 2015; Arora et al. 2018, 2020)
    - Export orientation (Blazquez & Domenech 2018)
    - Forecast stock market (Moat et al 2014)
    - Tourism demand (Alis et al. 2015)
    - Company survival and default prediction (Blazquez et al. 2018; Crosato et al. 2021)

Non-traditional sources of data

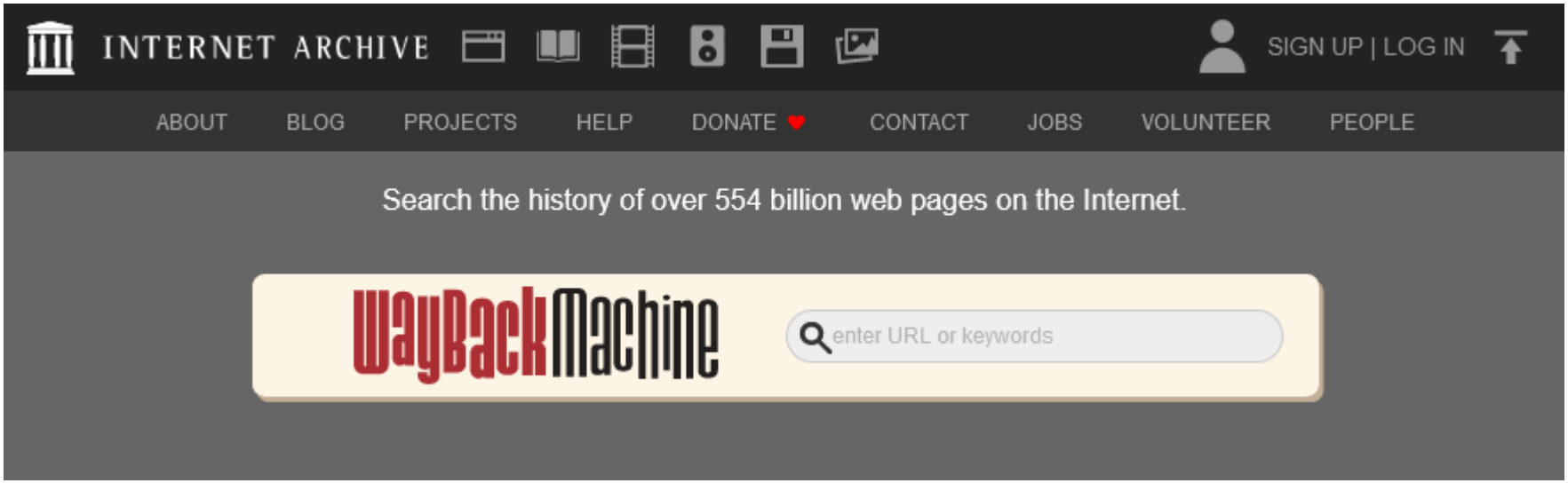
- Websites (II)
  - As opinion sharing platform:
    - Forecasting consumer product demand (Chong et al. 2015; Schneider & Gupta 2016)
    - Identification of influencers (Arenas-Márquez et al. 2014)
    - Tourist preferences (Li et al. 2015)
    - Track changes in the job search (Edelman, 2012)
  - As transaction platform:
    - Explain price differences between new and used products (Frota Neto et al. 2016)
    - E-commerce diffusion (Blazquez et al. 2019; Bruni & Bianchi 2020)

Non-traditional sources of data

- Sources within websites:
  - Page views and web analytics
  - Page content
    - Defining keywords and counting occurrences
    - Natural language processing
      - Language detection
      - Topic extraction
      - Sentiment analysis
  - HTML code
  - Server responses
  - Site-level analysis

Non-traditional sources of data

- Websites
  - Past versions are (partially) accessible with the **Wayback Machine** of the Internet Archive



2008

MEGA™
Creativity to the Rescue.™

This page requires Adobe Flash Player 8 or above and javascript enabled.

Click Here to download Flash Player

Home Page

About Us Shop Kid Zone Family Fun Customer Service

Contact Us
Legal Notice
Privacy Policy

Investor Relations
Investor Info
Financial Reports
Press Releases
Donations
Careers

Shop
- Construction Blocks
- Magnets
- Arts and Crafts
- Games and Activities
- Stationery
- Tech Toys

Award-Winners

Toys
- Magnetix
- Dragons
- Pirates Of The Caribbean 2
- Pirates Of The Caribbean 3
- Pyrates
- Plasmaverse
- Neoshifters
- B'Chic

Movie/Animation
- Magnetix
- Magnaman
- Magwarriors
- Dragons
- Dragons Metal Ages
- Pirates Of The Caribbean 3
- Spiderman 3

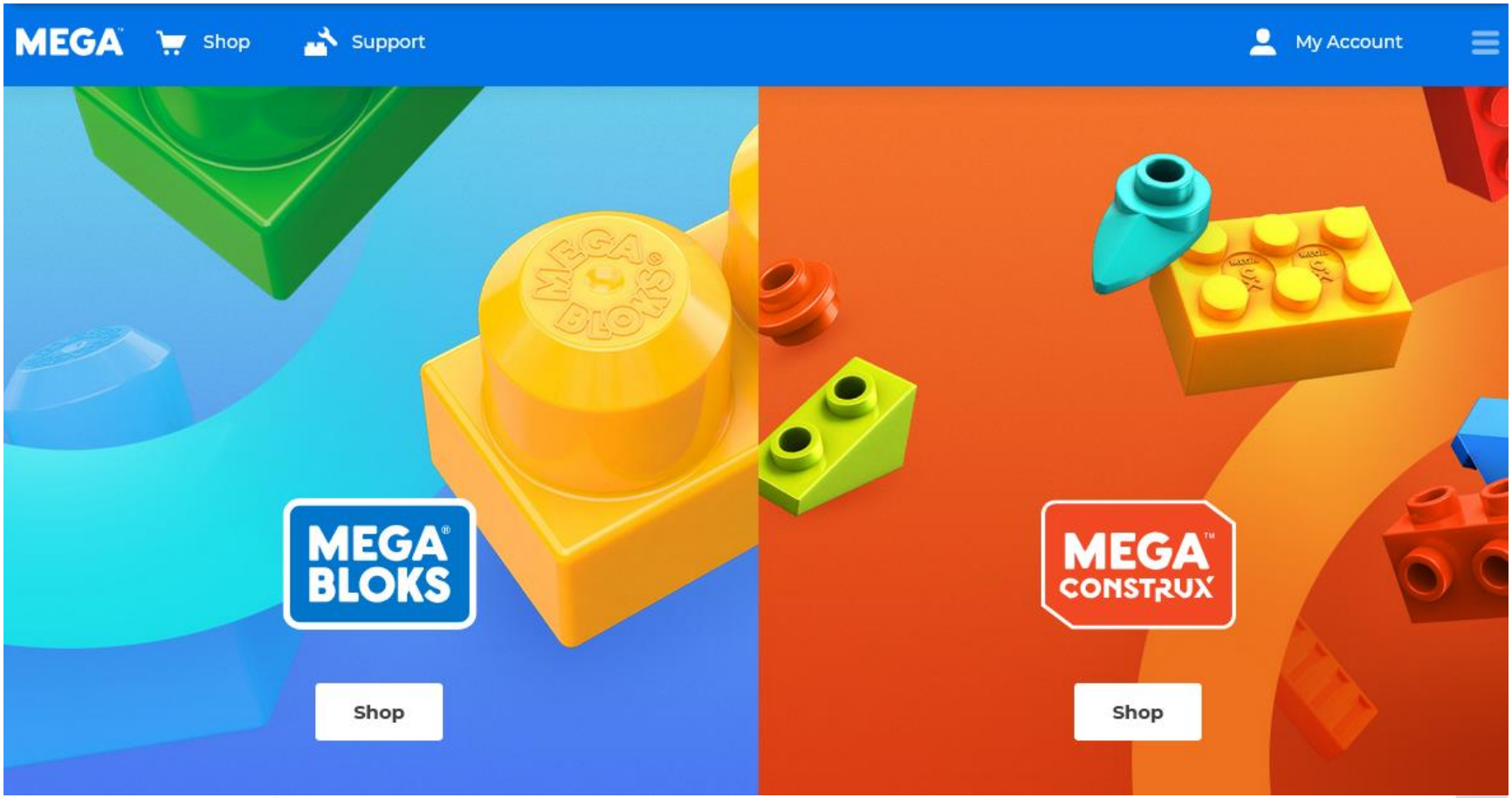
Family Fun
Activity Center
Learning Benefits
MEGA Club
Parent Panel
Contest

Request Center
Instructions
F.A.Q
Recall Information

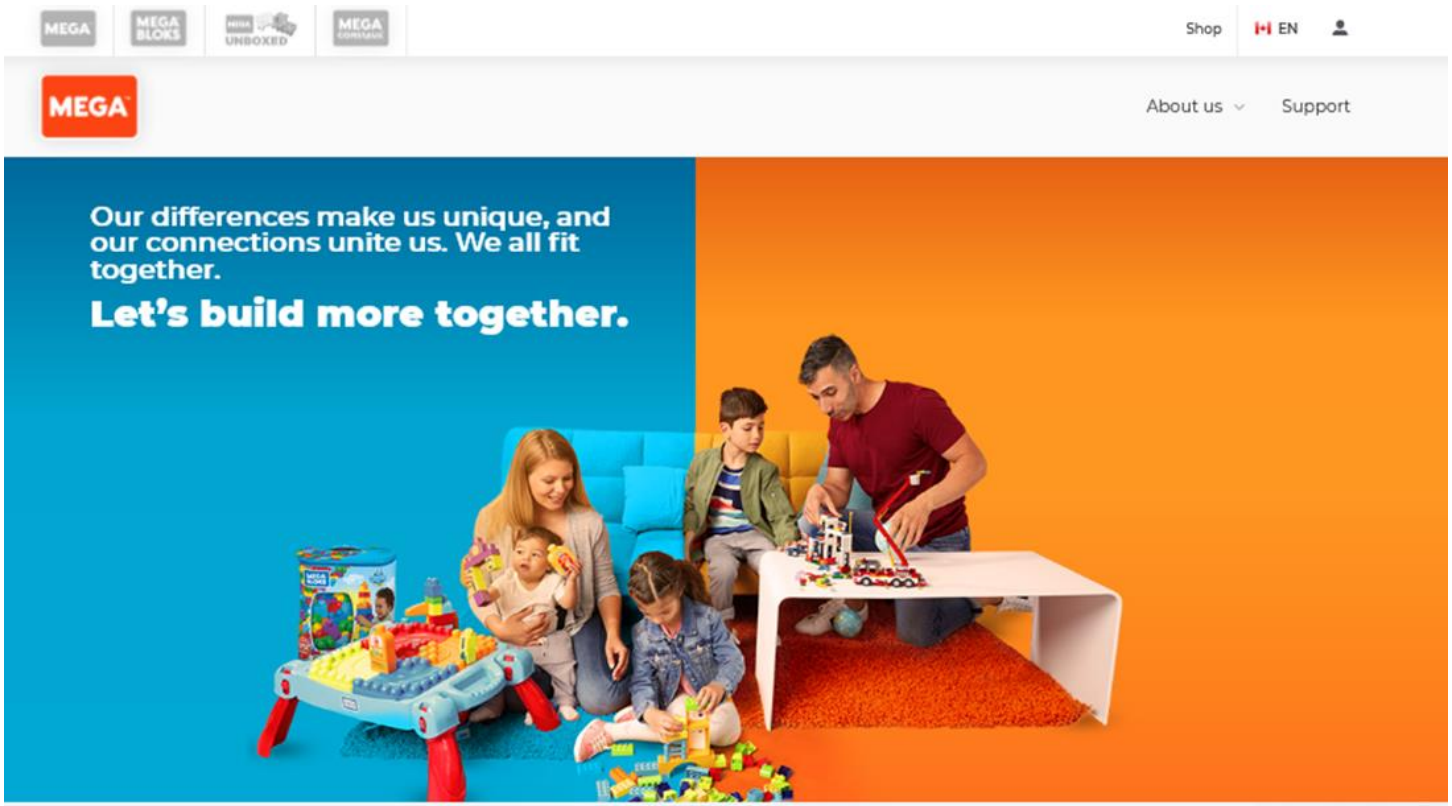
2011

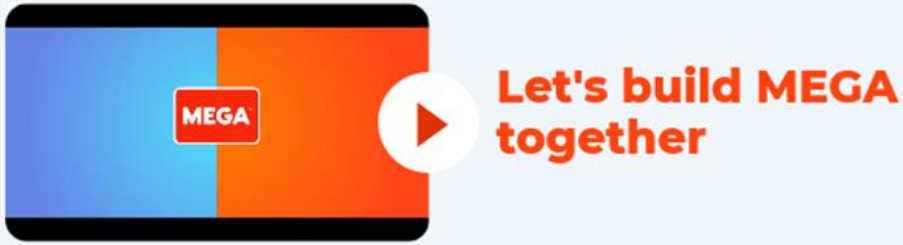

Français | Español
Search
MEGA
Home
Brands
Kids Zone
Corporate
Family Club
Videos
BLOK SQUAD
VISIT THE WEBSITE
MEGA BLOKS
RoseArt
MEGA PUZZLES
The BOARD DUDES

2019


MEGA
Shop
Support
My Account
MEGA BLOKS
Shop
MEGA CONSTRUX
Shop

2021


MEGA
MEGA BLOKS
MEGA UNBOXED
MEGA CONSTRUX
Shop
EN
MEGA
About us
Support
Our differences make us unique, and our connections unite us. We all fit together.
Let's build more together.


MEGA
Let's build MEGA together

Non-traditional sources of data

- Social Networking Sites
  - The information they contain is to some extent a reflection of what happens in the society
    - “Social Big Data”
  - Twitter: Most popular microblogging service. Research applications to:
    - Describe political preferences and forecast elections results (Tumasjan et al. 2011; Kim & Park 2012; Ceron et al. 2014)
    - Predict stock market movements (Bollen et al. 2011)
    - Forecast box office in film industry (Kim et al. 2015; Gaikar et al. 2015)
    - Monitor public opinion on new policies (Ceron & Negri 2016)

Non-traditional sources of data

- Social Networking Sites (II)
  - Facebook
    - Heterogeneous and user-adapted contents
      - It is more difficult to retrieve or analyze
    - Research applications to:
      - Determine consumer profiles (Arrigo et al. 2016)
      - Predict election results and political orientation (Cameron et al. 2016; David et al. 2016)
  - Other sources include LinkedIn, YouTube, Instagram, Tumblr, Flickr, blogs… (Bello-Orgaz et al. 2016)
  - SNS are commonly biased towards some segments of the population, so some correcting measures should be considered (Gayo-Avello 2012)

Non-traditional sources of data

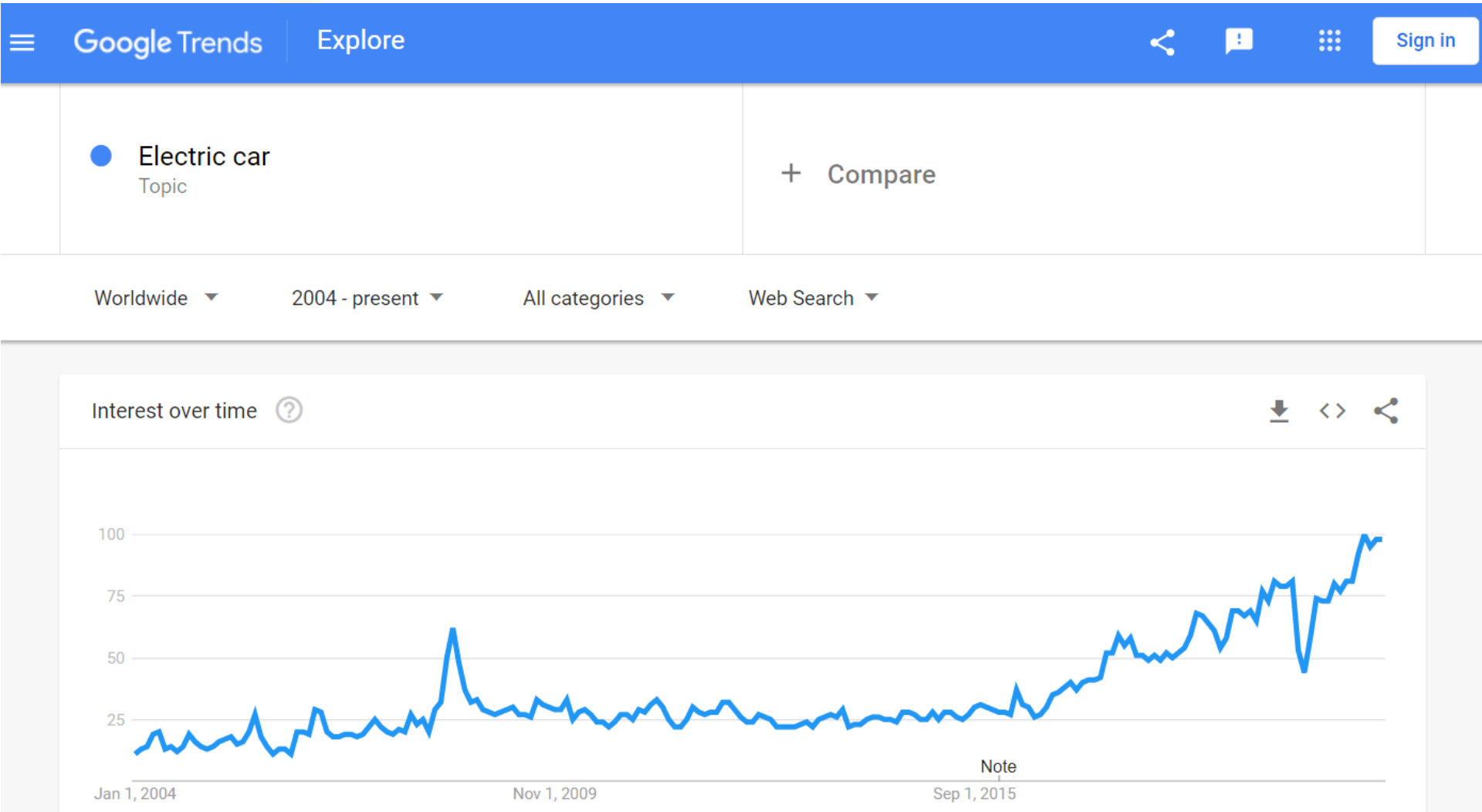
- Innovation in Twitter

Non-traditional sources of data

- Innovation in Instagram

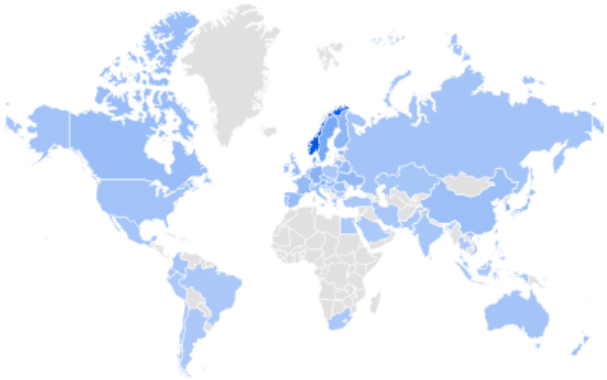
Non-traditional sources of data

- Search engines and Google Trends (GT)
  - Reports up-to-date data on the volume search of queries
    - It does not provide the absolute number of queries but a Search Volume Index (SVI)
    - Weekly data
    - Country / Region / City levels
  - Used in research since 2009:
    - Car and home sales, incoming tourists, unemployment (Choi & Varian 2009)
    - Trading decisions and transaction volumes on the stock market
    - Private purchases of goods and services (Vosen & Schmidt 2011)
    - Cinema admissions (Hand & Judge 2012)


Google Trends
Explore
Sign in
Electric car
Topic
Compare
Worldwide
2004 - present
All categories
Web Search
Interest over time
100
75
50
25
Note
Jan 1, 2004
Nov 1, 2009
Sep 1, 2015

Electric car — Worldwide, 2004 - present

Interest by region — Region



1	Norway	100
2	Taiwan	44
3	Denmark	36
4	Sweden	32
5	South Korea	29

Include low search volume regions

Showing 1-5 of 62 regions

Related topics — Rising

1	Tesla, Inc. - Electric car company	Breakout
2	Charging station - Topic	Breakout
3	Tesla - Automobile make	Breakout
4	Nissan - Automobile make	Breakout

Related queries — Rising

1	tesla	Breakout
2	電動 車	Breakout
3	elektrische auto	Breakout
4	tesla electric car	Breakout

Non-traditional sources of data

- Urban and mobile sensors (I):
  - Credit card reader
    - Personal bankruptcy detection (Xiong et al. 2013)
    - Fraudulent purchases (Van Vlasselaer et al. 2015)
    - Default and repayment (Einav & Levin 2014)
  - Retail scanners
    - Model market trends (Dey et al. 2014)
    - Detect consumer boycotts (Pandya & Venkatesan 2016)
  - Toll data
    - Economic cycle nowcasting (Askitas & Zimmerman 2013)

Non-traditional sources of data

- Urban and mobile sensors (II):
  - Call Detail Records and mobile phone data
    - Mobility patterns (Williams et al. 2015; Laurila et al. 2013)
    - Detection of places of interest (Montoliu et al 2013)
    - Analysis of population distribution (Deville et al. 2014; Graells-Garrido 2016)
  - Other sensors: tourist cards, WIFI access points (Kitchin 2014; Chou & Ngo 2016)

Non-traditional methods

- Non-traditional sources are traditionally large, heterogeneous and unstructured or semi-structured
  - New challenges regarding data management:
    - Retrieval
    - Processing
    - Analysis
    - Storage
  - Methods for dealing for such quantity of data have been applied in other fields, but they are relatively new in social and economic sciences (Varian 2014)

Non-traditional methods

- Classification of methods:
  - Structuring data
  - Modeling relations
  - Assessing models

Non-traditional methods

- Methods for structuring data
  - Data sources are traditionally classified as:
    - Structured (tabular data)
    - Semi-structured (includes some machine-readable tags)
    - Unstructured (without schemes to machine-interpret it)
  - Since most big data is not structured, a processes for transforming the data into an organized set is required

Non-traditional methods

- Methods for structuring data
  - Natural Language Processing (NLP)
    - It is a research area encompassing many techniques to make computers understand human-written text
    - Some techniques include:
      - Bag of Words
      - TF-IDF
      - Stemming and Lemmatization
      - Sentiment analysis (or Opinion Mining)
      - Latent Semantic Analysis
      - Latent Dirichlet Allocation
    - These techniques are more developed in English than in any other language

Non-traditional methods

- Methods for structuring data
  - Data matching
    - Technique for linking records from the same user (or entity) across different data sources
    - A special case is deduplication
      - Identification and matching of different records about the same entities in the same dataset

Non-traditional methods

- Methods for modeling
  Two main paradigms:
  - Supervised Learning:
    - Each observation in the dataset has inputs and outputs
      - Inputs is equivalent to independent variables, features or predictors
      - Output is equivalent to dependent variables, targets or responses
  - Unsupervised Learning:
    - Each observation has inputs but no outputs
    - The objective is to find the relationships or structure among inputs

Non-traditional methods

- Methods for modeling
  - Supervised Learning:
    - Supervised learning problems can be divided into:
      - Classification problems (discrete output)
      - Regression problems (continuous output)
    - Includes traditional methods such as linear and logistic regressions, but also:
      - Decision trees
      - Support Vector Machines
      - Artificial Neural Networks
      - Deep Learning
    - Nowcasting and forecasting applications generally use these methods

Non-traditional methods

- Methods for modeling
  - Unsupervised Learning:
    - Supervised learning problems can be divided into:
      - Clustering problems (discover groupings)
      - Association problems (find rules to describe the data)
    - Includes traditional methods such as PCA, but also:
      - Artificial Neural Networks
      - Deep Learning

Non-traditional methods

- Methods for modeling
  - There are some other techniques for improving the performance of the previous methods
    - Ensemble algorithms, such as bootstrap, random forests…
    - Regularization methods, such as LASSO, elastic net, ridge regression…
    - Bayesian methods, such as model averaging, BSTS, Spike-and-Slab regression…

Non-traditional methods

- Methods for assessing models
  - The objective is to obtain a **robust** model with the best out-of-sample predictive **performance**
    - Performance refers to how well it fits the data
    - Robust refers to how well it works with other data
  - Traditional tests (R^2, AIC, BIC, Log-likelihood…) were not conceived for treating a huge amount of complex data (Varian 2014)

Non-traditional methods

- Methods for assessing models
  - A holdout process helps to ensure the robustness. The initial sample is split into:
    - Training set
    - (Validation set)
    - Test set
  - Approaches:
    - K-fold Cross-Validation
    - Leave-one-out Cross-Validation

Non-traditional methods

- Methods for assessing models
  - To properly evaluate classifiers:
    - The train set should be balanced
      - Solutions such as oversampling, undersampling or SMOTE could be applied
    - Predictive accuracy analysis with:
      - Precision-Recall curves
      - ROC curves
      - Confusion matrix
    - Cost-sensitive analysis:
      - Applies a Cost Matrix reflecting the costs associated to each misclassification

Conclusion

- Digital data comes with many opportunities, but more challenges:
  - Change of data management systems
  - New analysis methods
  - Need for multidisciplinary approaches

Non-traditional data sources for social and economic monitoring

Josep Domenech
Universitat Politècnica de València

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*. 10.1016/j.techfore.2017.07.027