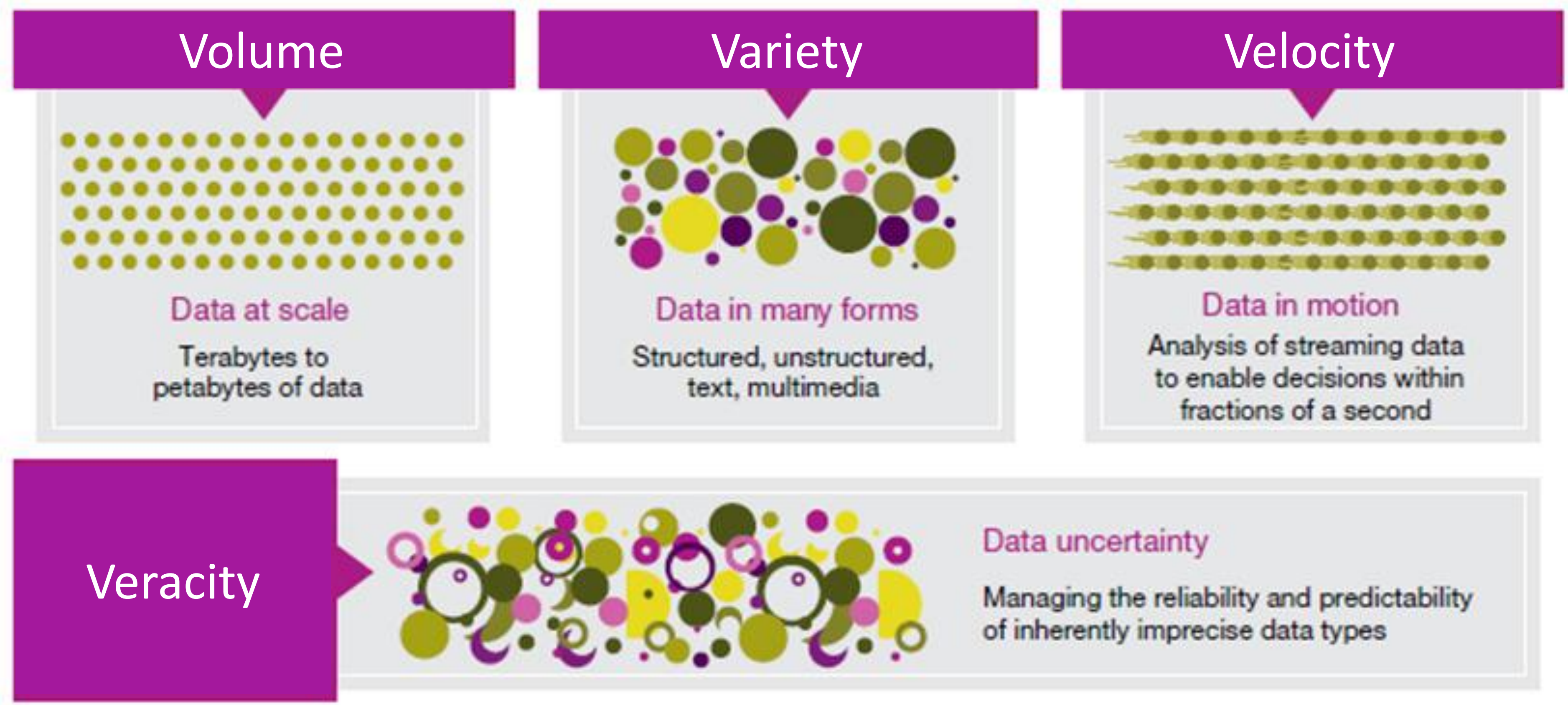
Population and housing census and Big Data

March 3, 2015

INEGI

INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Big data dimensions



Volume

Data at scale

Terabytes to petabytes of data

Variety

Data in many forms

Structured, unstructured, text, multimedia

Velocity

Data in motion

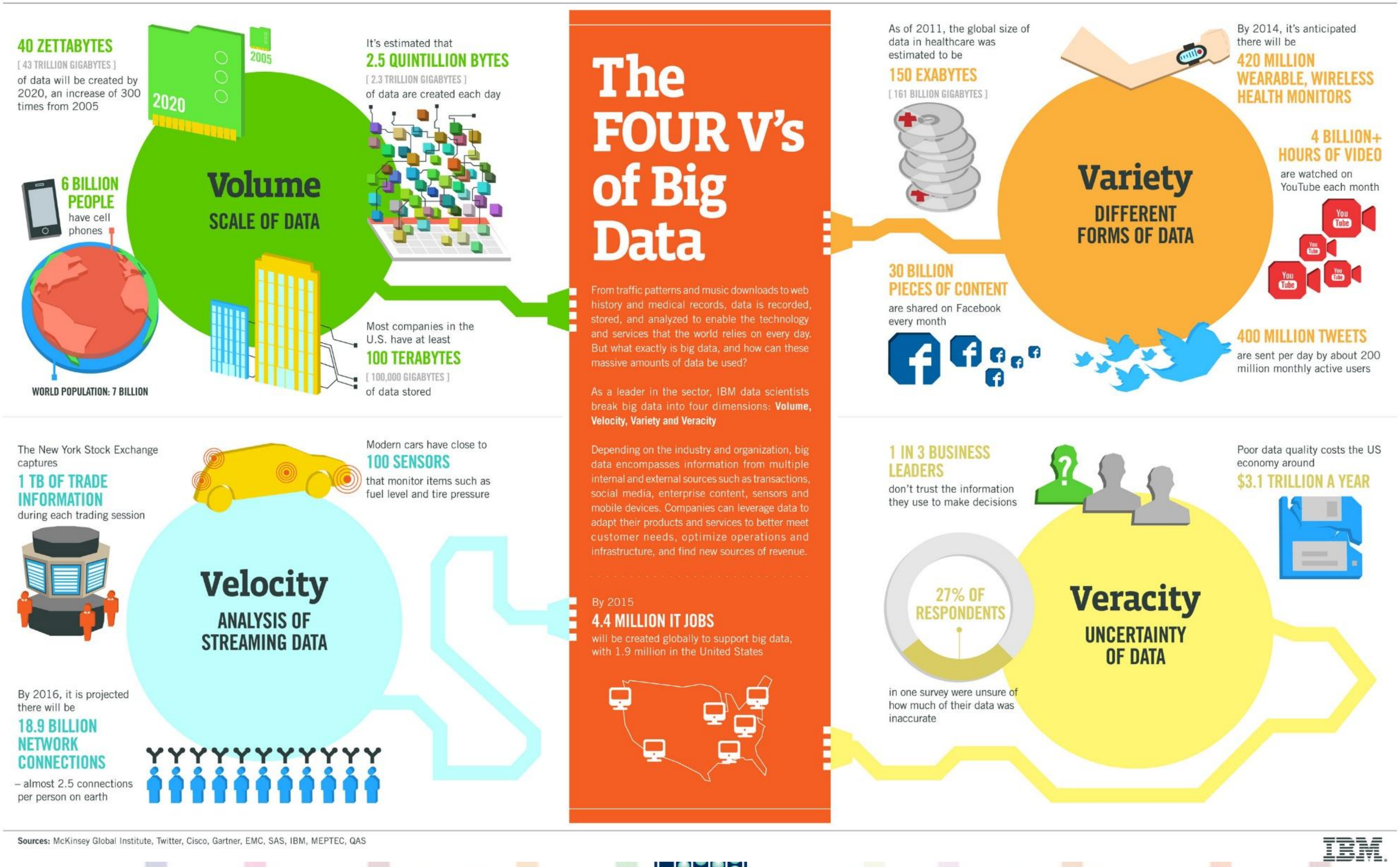
Analysis of streaming data to enable decisions within fractions of a second

Veracity

Data uncertainty

Managing the reliability and predictability of inherently imprecise data types

Big data dimensions



The FOUR V's of Big Data

From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?

As a leader in the sector, IBM data scientists break big data into four dimensions: **Volume, Velocity, Variety and Veracity**

Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.

By 2015 **4.4 MILLION IT JOBS** will be created globally to support big data, with 1.9 million in the United States

Volume
SCALE OF DATA

40 ZETTABYTES [43 TRILLION GIGABYTES] of data will be created by 2020, an increase of 300 times from 2005

It's estimated that **2.5 QUINTILLION BYTES** [2.3 TRILLION GIGABYTES] of data are created each day

6 BILLION PEOPLE have cell phones

WORLD POPULATION: 7 BILLION

Most companies in the U.S. have at least **100 TERABYTES** [100,000 GIGABYTES] of data stored

Velocity
ANALYSIS OF STREAMING DATA

The New York Stock Exchange captures **1 TB OF TRADE INFORMATION** during each trading session

Modern cars have close to **100 SENSORS** that monitor items such as fuel level and tire pressure

By 2016, it is projected there will be **18.9 BILLION NETWORK CONNECTIONS** – almost 2.5 connections per person on earth

Variety
DIFFERENT FORMS OF DATA

As of 2011, the global size of data in healthcare was estimated to be **150 EXABYTES** [161 BILLION GIGABYTES]

By 2014, it's anticipated there will be **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

Veracity
UNCERTAINTY OF DATA

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**

27% OF RESPONDENTS in one survey were unsure of how much of their data was inaccurate

Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS

Corporate big data

Bankcards registry

For 2012 in Mexico, users of formal credit were 27.5% of the adult population (19.3 million). Department store cards are the product most used (72.2%), followed by bank credit cards (32.9%). On the other hand, approximately 15 million adult have a debit card.

Registry of real estate purchases

According to the firm Investment Properties Mexico, in 2014, 95% of all real estate purchases began online.

The Mexican Association of Professional Realtors representing over 2,800 construction companies and near 30,000 realtors.

Corporate big data

Registration of commercial establishments

For 2014 in Mexico were registered 5,223 establishments classified as supermarkets that attend an average of 18 million customers per day.

DENUE 2014 and ANTAD

In Mexico there are around 14,000 establishments called convenience stores (mini markets). According to the company Kantar Worldpanel, 6 out of 10 households shop on this type of establishment and the average value of each purchase is 4 USD.

Corporate big data

19. 5 million dwellings (69%) have tubed water inside the house.

CPV 2010

5.8 million subscribers at TV-CABLE

MODUTIH 2012

27.5 million dwellings (98%) with electric energy

CPV 2010

Corporate big data

12.2 million dwellings (43.5%) with landline and 18.3 million (65.5%) have cell phone.

CPV 2010

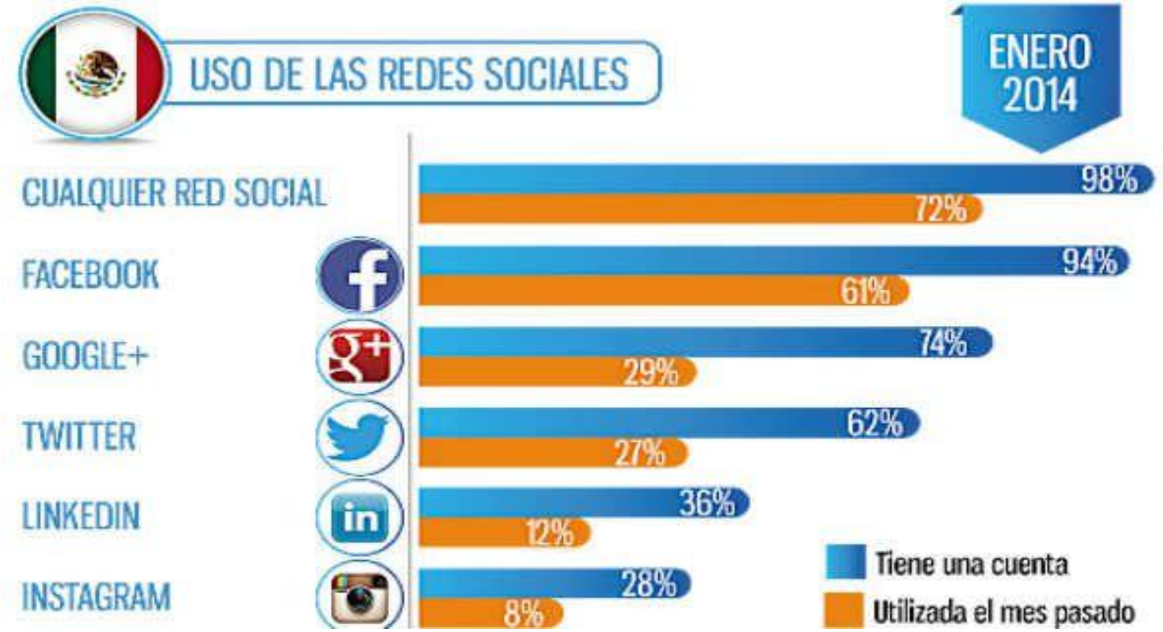
23.7 million dwellings (84%) use gas for cooking.

CPV 2010

30.7 million households have a gas tank.

ENGASTO 2013

Social networks big data



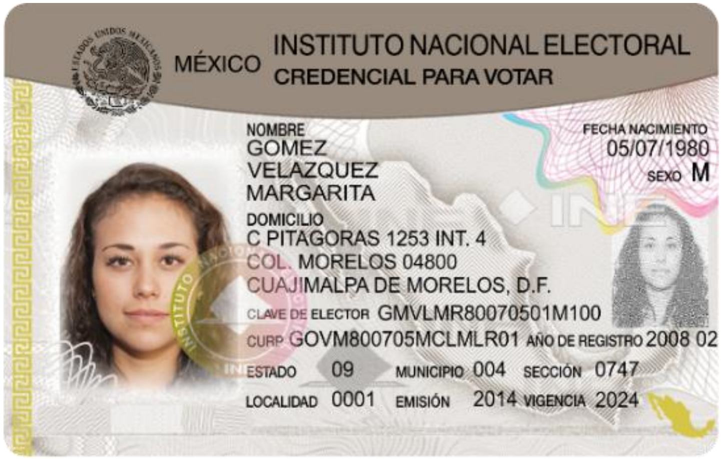
The Competitive Intelligence Unit (CIU).

In Mexico there are 47.4 million Internet users. 67.4% use Internet to look for information, 38.5 to communicate with others, **39.6% using social networks** and 1.5% for financial transactions or e-commerce transactions.

MODUTIH 2014

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

Governmental big data



MÉXICO INSTITUTO NACIONAL ELECTORAL
CREDENCIAL PARA VOTAR

NOMBRE
GOMEZ
VELAZQUEZ
MARGARITA

FECHA NACIMIENTO 05/07/1980
SEXO M

DOMICILIO
C PITAGORAS 1253 INT. 4
COL. MORELOS 04800
CUAJIMALPA DE MORELOS, D.F.

CLAVE DE ELECTOR GMVLMR80070501M100
CURP GOVM800705MCLMLR01 AÑO DE REGISTRO 2008 02
ESTADO 09 MUNICIPIO 004 SECCIÓN 0747
LOCALIDAD 0001 EMISIÓN 2014 VIGENCIA 2024

National Electoral Institute
In December 2014, Mexico there were 81.2 million people over 18 years that have the voting card.

INE, 2014

Pensions system

22.5 million people over 16 have a savings account for retirement.

ENGASTO 2013

INEGI
INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA

Governmental big data

Health care system
71.6 million people are beneficiaries to health services in public institutions

CPV 2010

Urban public transport system
In December 2014, only in Mexico City were registered 22.8 million of passengers to the different public transport services: bus, metrobus and subway

DF government

Governmental big data

Registers of Foreign Ministry

In 2013 were issued 3,581 naturalization certificates.

Tax collection system

In March 2014, the number of registered taxpayers in the Tax Administration Service (SAT) amounted to 42.8 million people.

Governmental big data

Land registry

State governments have data on the number, size and value of the plots of land registered in each of the 2,457 municipalities of Mexico (100%).

Public registry of property

14. 7 million of households

have deed title (47.9%)

ENGASTO 2013

Census of population and housing of Mexico

Demographic variables

Culture variables

Housing variables

Social variables

Economic variables

Household variables

Main focus in terms of exploiting big data for the purpose of the population and housing census in the 2020 Round?

Growth of housing registered in the Register of Housing (RUV), after the 2010 census.

1’299,045 dwellings updated

Main focus in terms of exploiting big data for the purpose of the population and housing census in the 2020 Round?

Births Information System

In 2012, 1’901,394 children were born.

Educational system

In the scholar year 2013-2014 had registered:

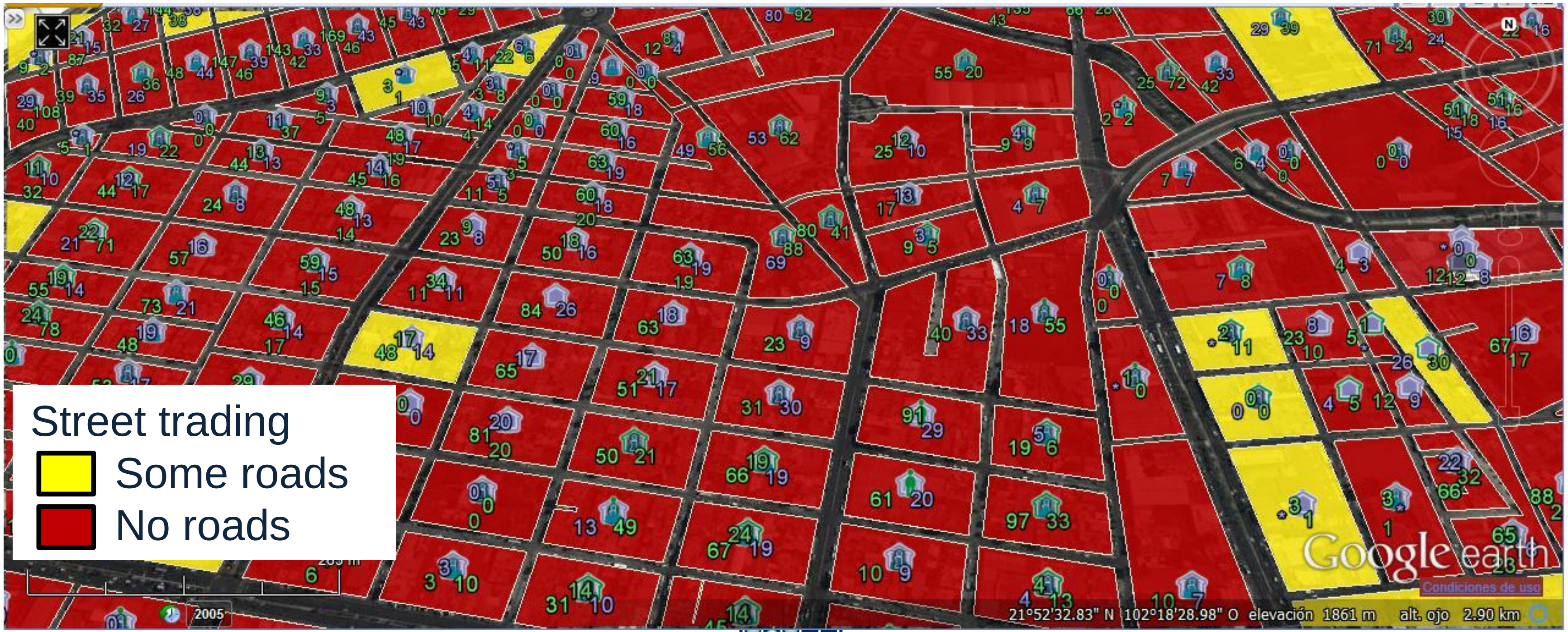
Total	35.7 million
Basic education	25.9 million
High school education	4.7 million
Higher education	3.4 million
Job training	1.7 million

Main focus in terms of exploiting big data for the purpose of the population and housing census in the 2020 Round?

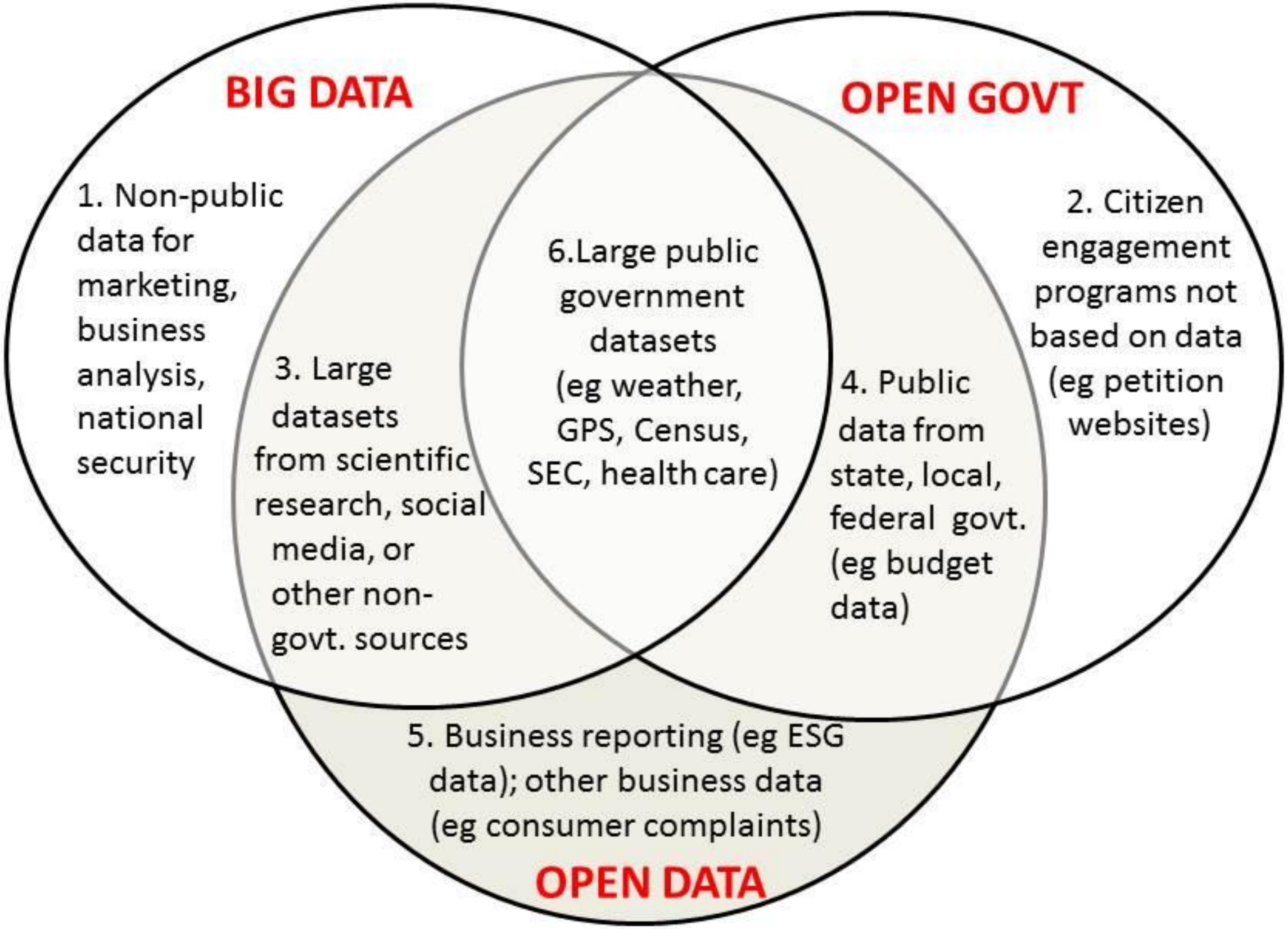
National Inventory of Housing

New platforms for integrating census and survey data with Big Data.

- Population of census
- Streets vitalities
- Services in the housing
- Street trading


Street trading
🟨 Some roads
🟥 No roads
Google earth
2005
21°52'32.83" N 102°18'28.98" O elevación 1861 m alt. ojo 2.90 km

Is the primary focus on governmental big data, corporate big data or both?



BIG DATA

OPEN GOVT

OPEN DATA

1. Non-public data for marketing, business analysis, national security

2. Citizen engagement programs not based on data (eg petition websites)

3. Large datasets from scientific research, social media, or other non-govt. sources

4. Public data from state, local, federal govt. (eg budget data)

5. Business reporting (eg ESG data); other business data (eg consumer complaints)

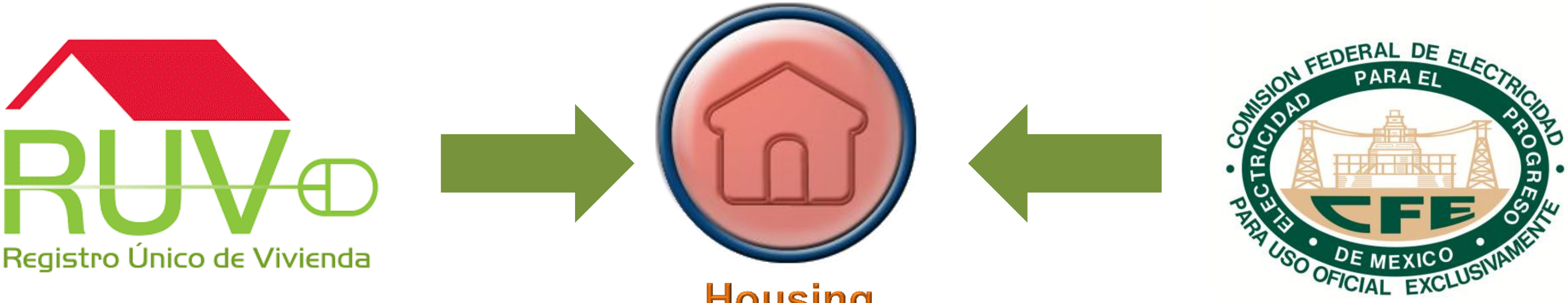
6.Large public government datasets (eg weather, GPS, Census, SEC, health care)

What are the main findings so far?

"What we really need in IT
is someone who has super powers."

What are the main findings so far?

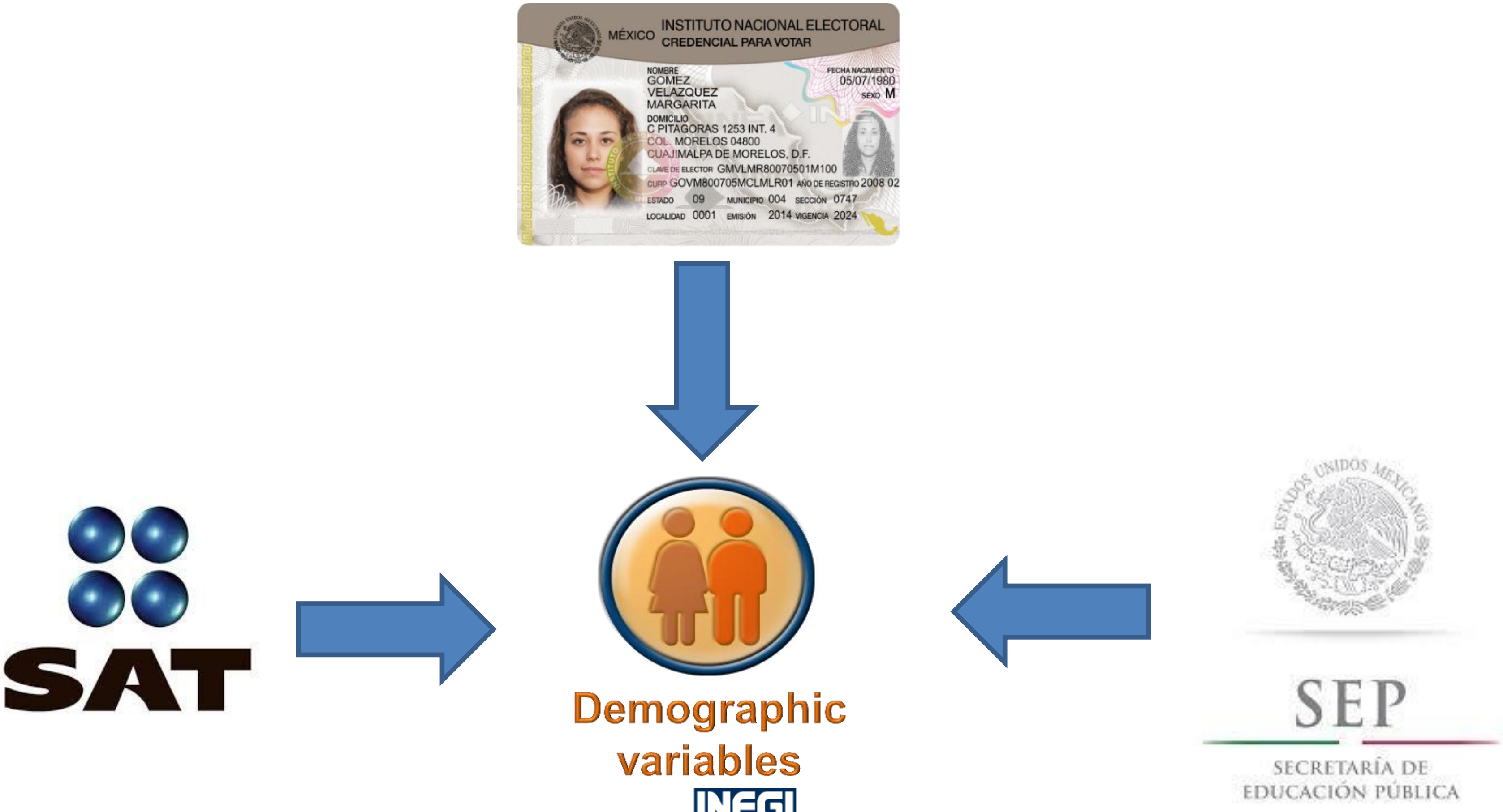
Update measurement of dwellings


RUV
Registro Único de Vivienda
COMISION FEDERAL DE ELECTRICIDAD
PARA EL PROGRESO
CFE
DE MEXICO
PARA USO OFICIAL EXCLUSIVAMENTE

Housing variables

What are the main findings so far?

Update measurement of Mexican population



MÉXICO INSTITUTO NACIONAL ELECTORAL
CREDENCIAL PARA VOTAR

NOMBRE
GOMEZ
VELAZQUEZ
MARGARITA

FECHA NACIMIENTO 05/07/1980
SEXO M

DOMICILIO
C PITAGORAS 1253 INT. 4
COL. MORELOS 04800
CUAJIMALPA DE MORELOS, D.F.

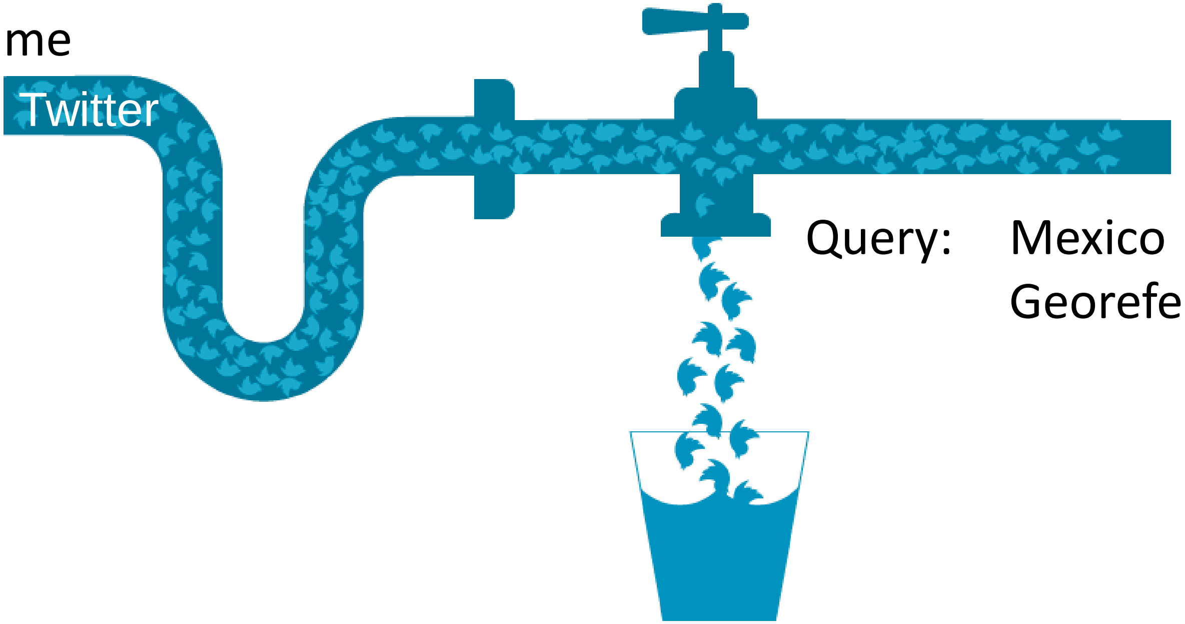
CLAVE DE ELECTOR GMVLMR80070501M100
CURP GOVM800705MCLMLR01 AÑO DE REGISTRO 2008 02
ESTADO 09 MUNICIPIO 004 SECCIÓN 0747
LOCALIDAD 0001 EMISIÓN 2014 VIGENCIA 2024

INEGI's explorations in Big Data

INEGI
INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

What are the main findings so far?

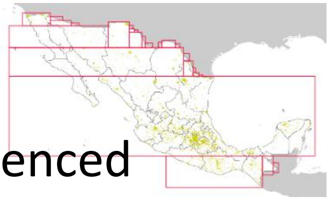
Twitter as a data source



Real time

Twitter

Query: Mexico Georeferenced



NoSQL Data Base

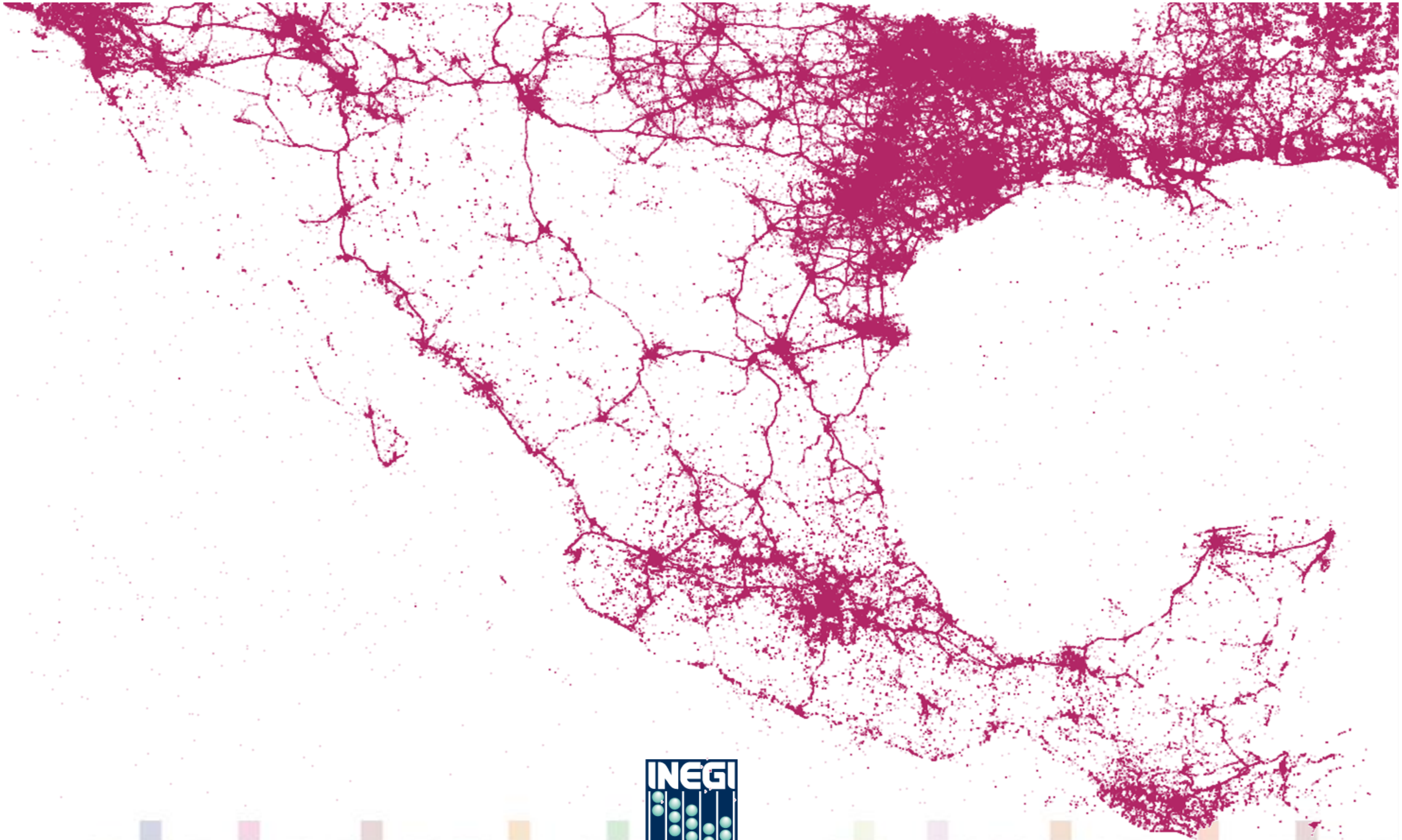
What are the main findings so far?

Why Twitter?

- Readily available
- Up to 1% of global tweets at no cost
- Around 12 M accounts in Mexico
- Geo located tweets by 700 thousand accounts
- 110 M plus tweets downloaded since January 2014
- Even though its drawbacks: Not documented, not supported by “traditional” statistics methodologies

What are the main findings so far?

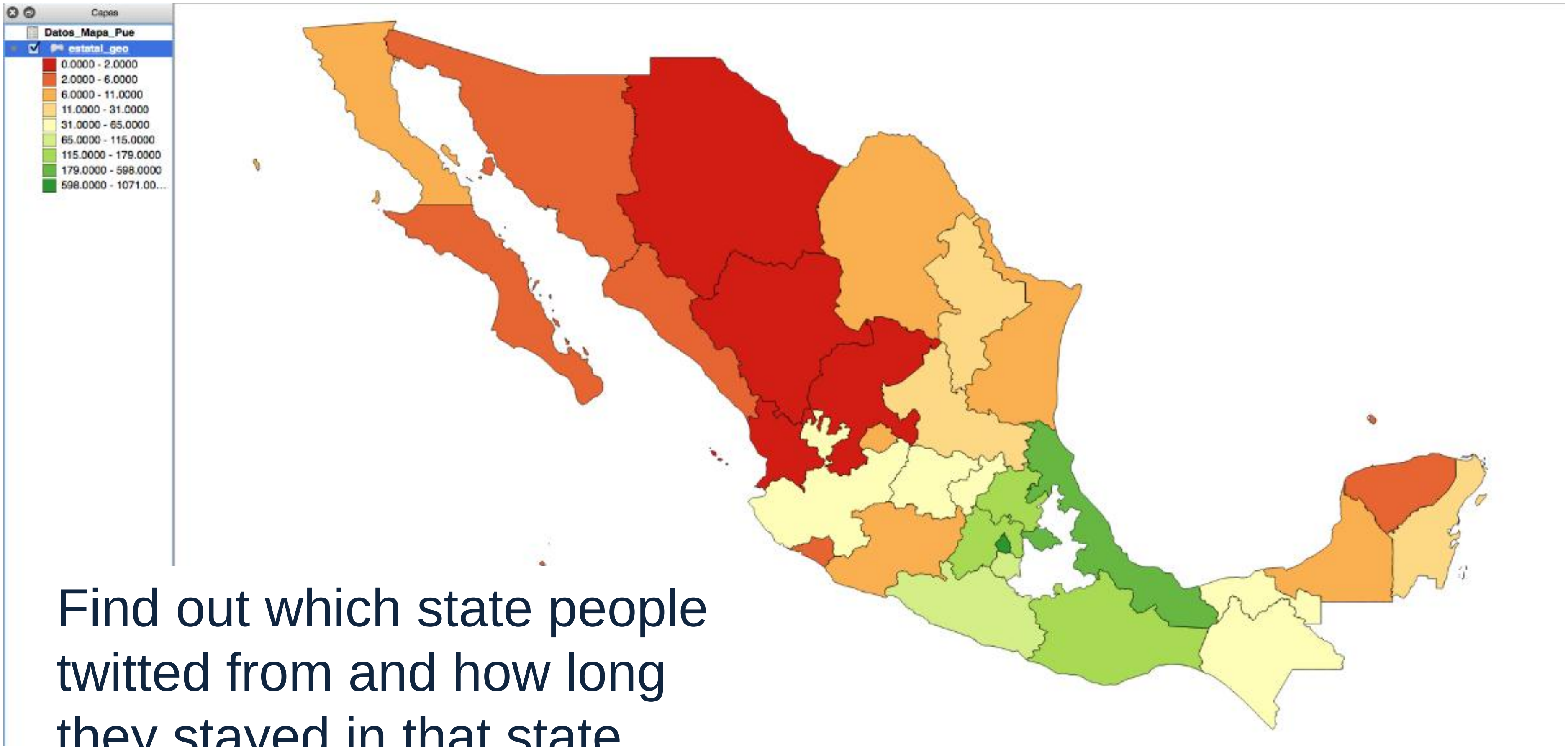
How do tweets look like?



INSTITUTO NACIONAL
DE ESTADÍSTICA Y GEOGRAFÍA

What are the main findings so far?

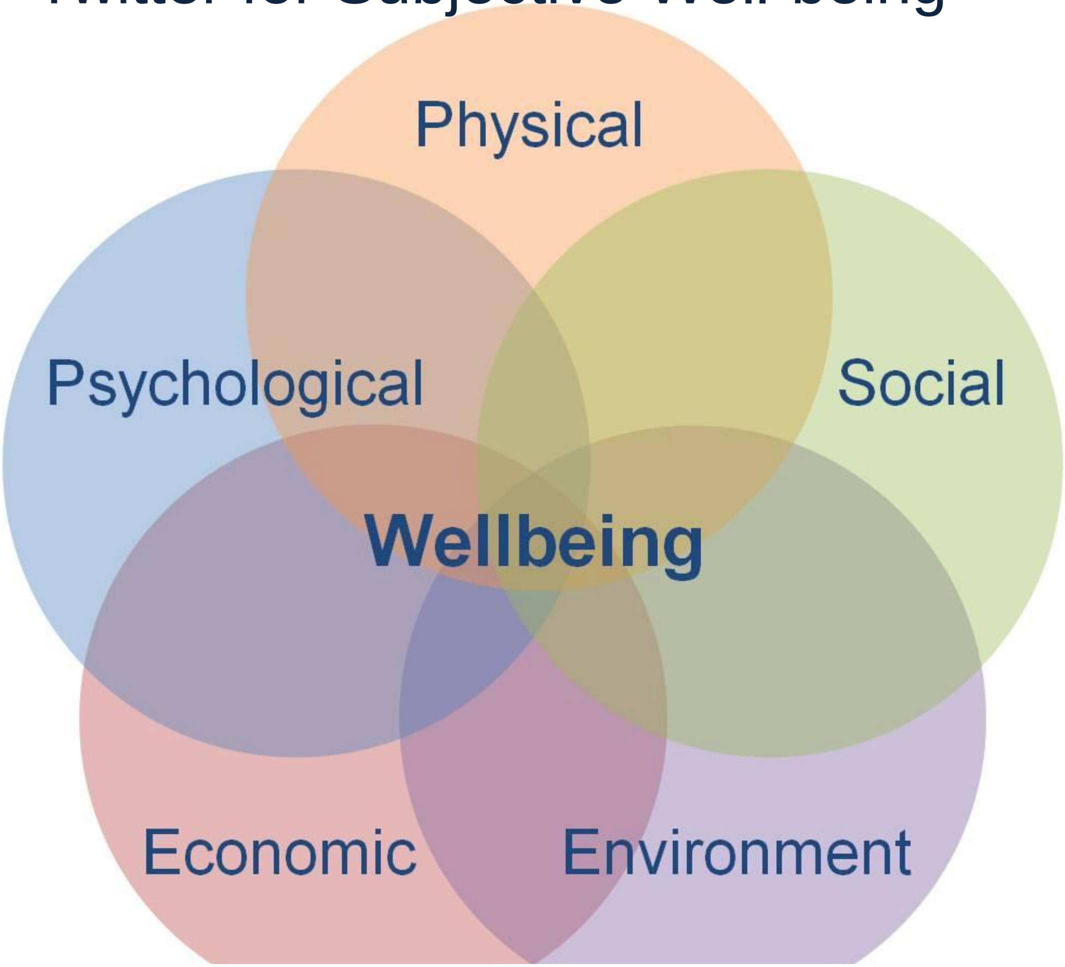
Tourists visiting Puebla during February 1 to 3



Find out which state people twitted from and how long they stayed in that state.

What are the main findings so far?

Twitter for Subjective Well-being



Humans grade a sample of tweets as positive, neutral or negative, and classify them in several subjects.

What are the challenges?

Analyzing Big Data

- Evaluation of its characteristics
  - How many registers?

  - There are special access conditions?
    - Income or expenditure levels
    - Educational level
    - Others

What are the challenges?

Analyzing Big Data

- Diagnosis of their potential and complementarity with census data

  - Definitions of the target population

  - Conditions for registration

What are the challenges?

Analyzing Big Data

- Technical capacity building
    - Integration of specialized teams
    - Agreements with universities and research centers for the use of data in combination
    - Promotion of training workshops

Some questions

- How will the scientific community and policy makers react to official statistics, especially small-area estimates, that rely on Big Data?

- What will be the public perception of these estimates?
    - Possible implications of a backlash.

- Will we be able to verify the veracity of Big Data?

Some questions

- Will the costs associated with Big Data prove to be unacceptable relative to the benefits? Consider a big implementation requires.
  - Software and hardware cost
  - Cost for development of technical capacities
  - Useful life time

Some questions

- Do we understand and can we describe what Big Data represent?
    - Cell phone users
    - Credit cars users