

# DDAMINTRODUCTION TO BIGDATA

Docente: Patrizio Dazzi

Mail: [patrizio.dazzi@isti.cnr.it](mailto:patrizio.dazzi@isti.cnr.it)

2005



Luca Bruno / AP

# DIGITAL LITTLE THUMBLINGS



Every year, each person leaves behind more than 5 GB of digital breadcrumbs.





1 NEW DEFINITION IS ADDED ON UPDICTIONARY

1,600+ READS ON Scribd

13,000+ HOURS MUSIC STREAMING ON PANDORA

12,000+ NEW ADS POSTED ON craigslist

370,000+ MINUTES VOICE CALLS ON skype

98,000+ TWEETS



320+ NEW twitter ACCOUNTS



100+ NEW Linked in ACCOUNTS

THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!

1 NEW ARTICLE IS PUBLISHED

6,600+ NEW PICTURES ARE UPLOADED ON flickr



50+ WORDPRESS DOWNLOADS

695,000+ facebook STATUS UPDATES



125+ PLUGIN DOWNLOADS

79,364 WALL POSTS

510,040 COMMENTS



694,445 SEARCH QUERIES

Google

Google Search

1,700+ Firefox DOWNLOADS



60+ NEW BLOGS

1,500+ BLOG POSTS



70+ DOMAINS REGISTERED



600+ NEW VIDEOS

100+ Answers.com 40+ YAHOO! ANSWERS

QUESTIONS ASKED ON THE INTERNET...

25+ HOURS TOTAL DURATION



# VOLUME, VELOCITY, VARIETY, VERACITY

**40 ZETTABYTES**  
 (43 TRILLION GIGABYTES)  
 of data will be created by 2020, an increase of 300 times from 2005



## Volume SCALE OF DATA



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



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

## 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



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



**30 BILLION PIECES OF CONTENT** are shared on Facebook every month



## Variety DIFFERENT FORMS OF DATA



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



**400 MILLION TWEETS** are sent per day by about 200 million monthly active users



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



## Velocity ANALYSIS OF STREAMING DATA



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



**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**

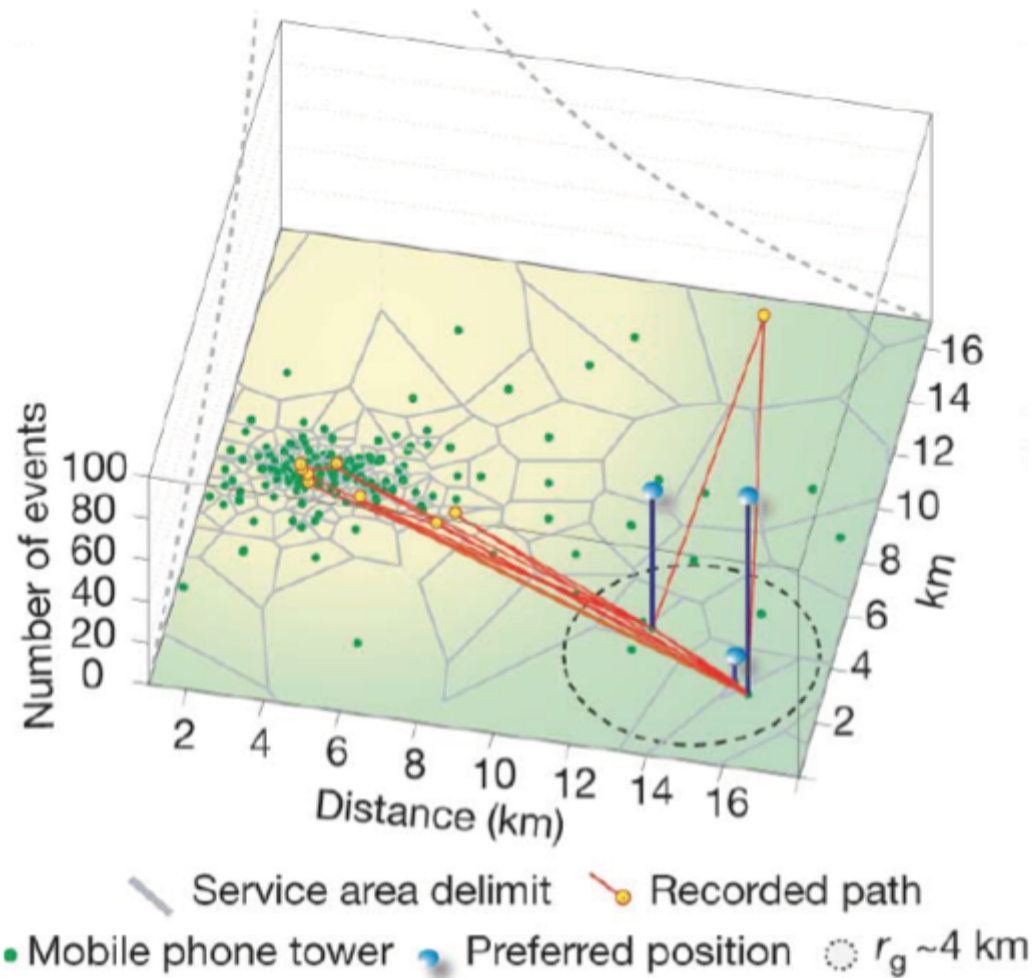


**27% OF RESPONDENTS**

in one survey were unsure of how much of their data was inaccurate

## Veracity UNCERTAINTY OF DATA

# CDR DATA



**when**  
you  
call



**where**  
you  
call



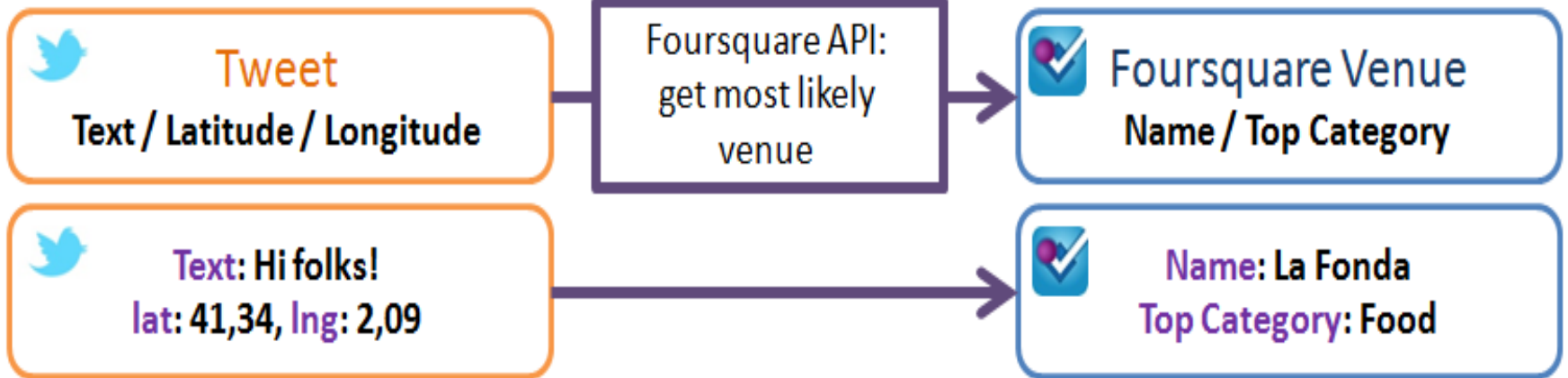
**who**  
you  
call



# SOCIAL NETWORKS

The image shows a screenshot of the Flickr website's geotagging interface. At the top, the Flickr logo is visible, along with navigation links for Home, The Tour, Sign Up, Explore, and Upload. A search bar is located in the top right corner. The main area is a map of Pisa, Italy, with several pink geotag markers placed on the map. A photo of the Leaning Tower of Pisa is displayed in a central window, with the caption "Pisa! by smalex.b". Below the map, a search bar is present with the text "34,639 geotagged items" and "Sort by: Interesting • Recent". The bottom left corner shows a scale bar (1250m) and the text "Data ©2010 NAVTEQ".

# TWITTER







# GPS VEHICLE TRACKS

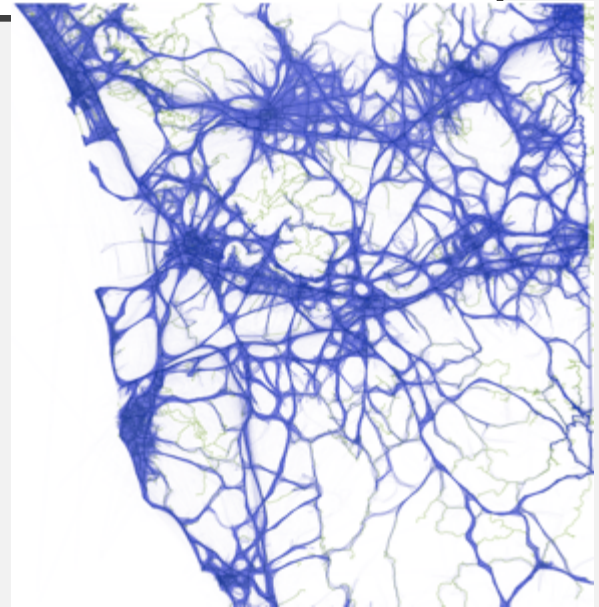
GPS fixes collected/sent by onboard navigation devices & “black boxes”

**Id;Time;Lat;Lon;Height;Course;Speed;PDOP;State;NSat**

```
...
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
8;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
8;22/03/07 08:51:59;50.777415;7.205543; 68.3;112.7;25.298;3.8;1808;4
8;22/03/07 08:52:03;50.777317;7.205877; 68.8;119.8;32.447;3.8;1808;4
8;22/03/07 08:52:06;50.777185;7.206202; 68.1;124.1;30.058;3.8;1808;4
8;22/03/07 08:52:09;50.777057;7.206522; 67.9;117.7;34.003;3.8;1808;4
8;22/03/07 08:52:12;50.776925;7.206858; 66.9;117.5;37.151;3.8;1808;4
8;22/03/07 08:52:15;50.776813;7.207263; 67.0;99.2;39.188;3.8;1808;4
8;22/03/07 08:52:18;50.776780;7.207745; 68.8;90.6;41.170;3.8;1808;4
8;22/03/07 08:52:21;50.776803;7.208262; 71.1;82.0;35.058;3.8;1808;4
8;22/03/07 08:52:24;50.776832;7.208682; 68.6;117.1;11.371;3.8;1808;4
```

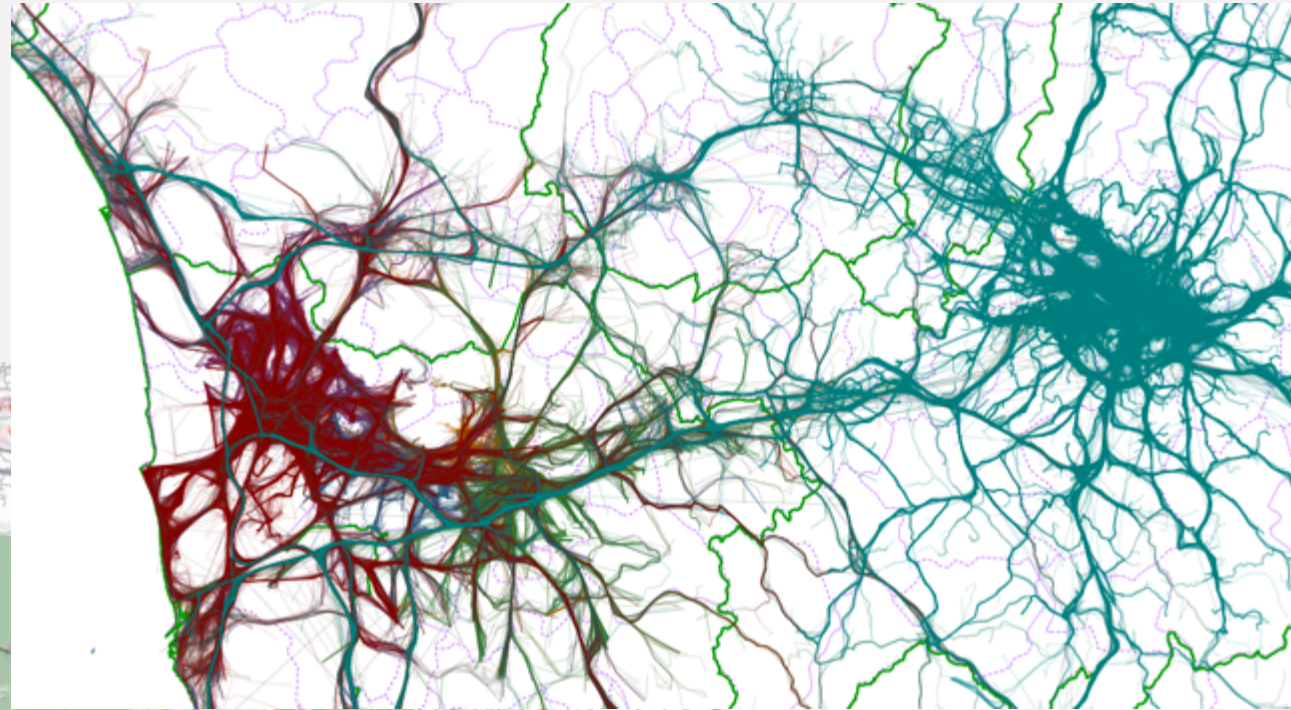
Typical sampling rate: 30-60 secs

Typical localization error: 5-10 m

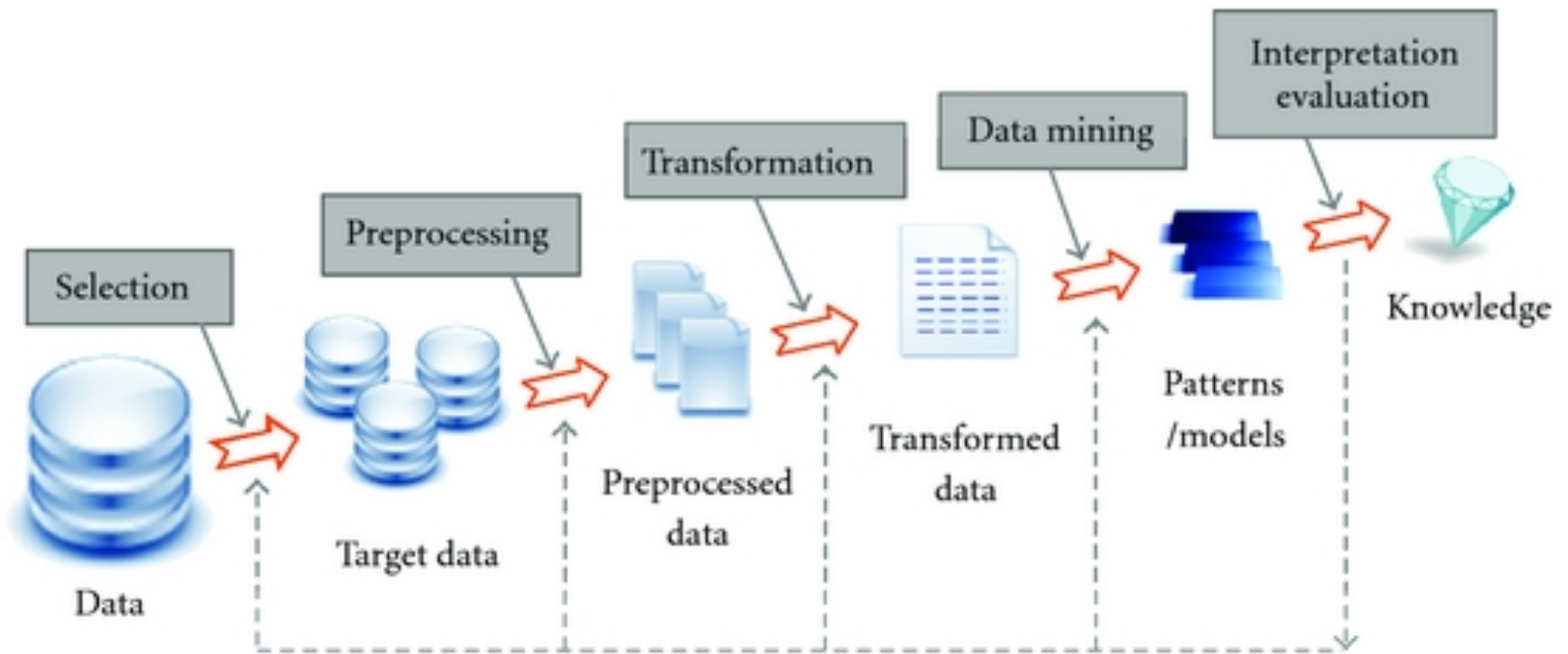




# GPS vehicle tracks

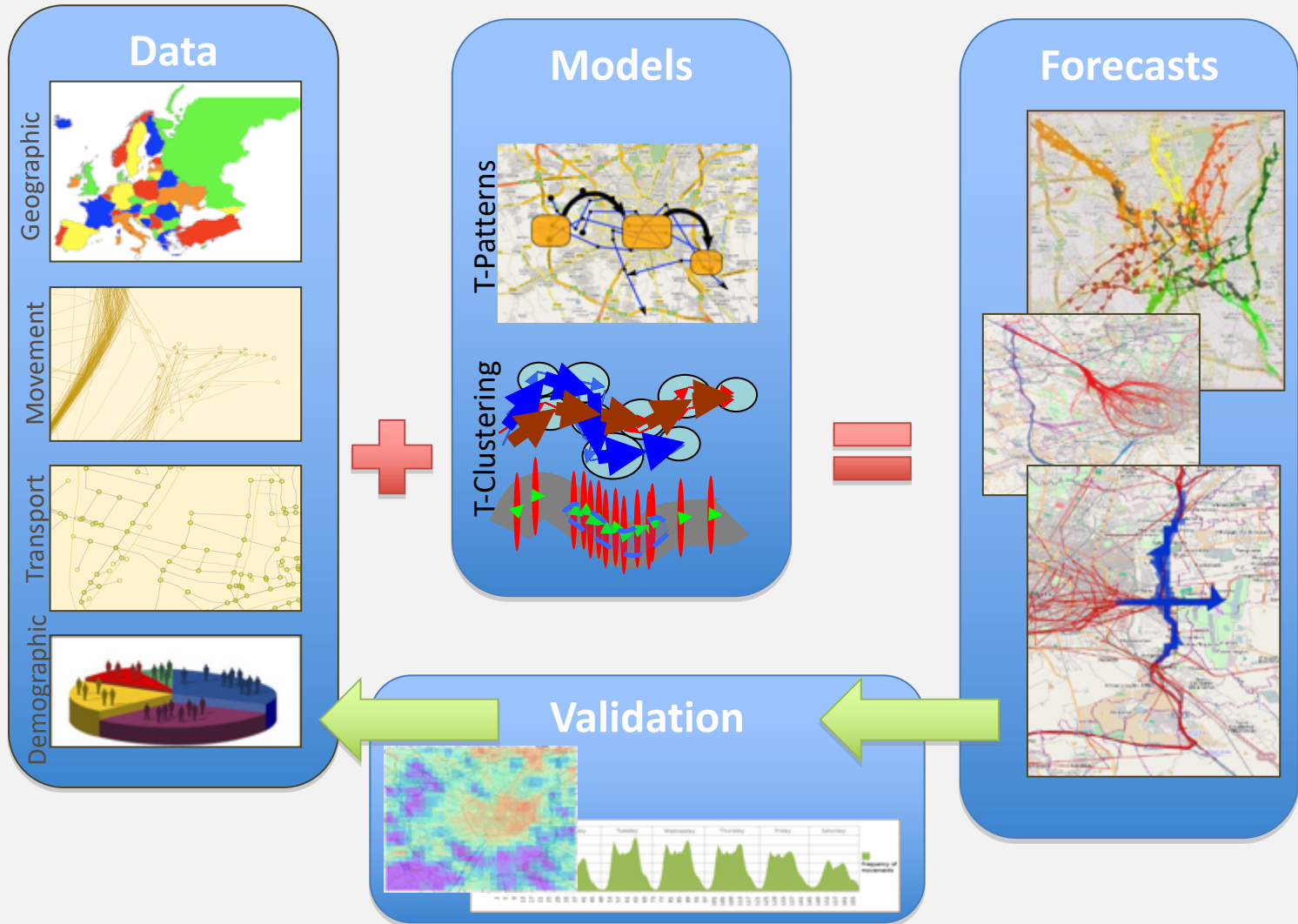


# THE KDD PROCESS





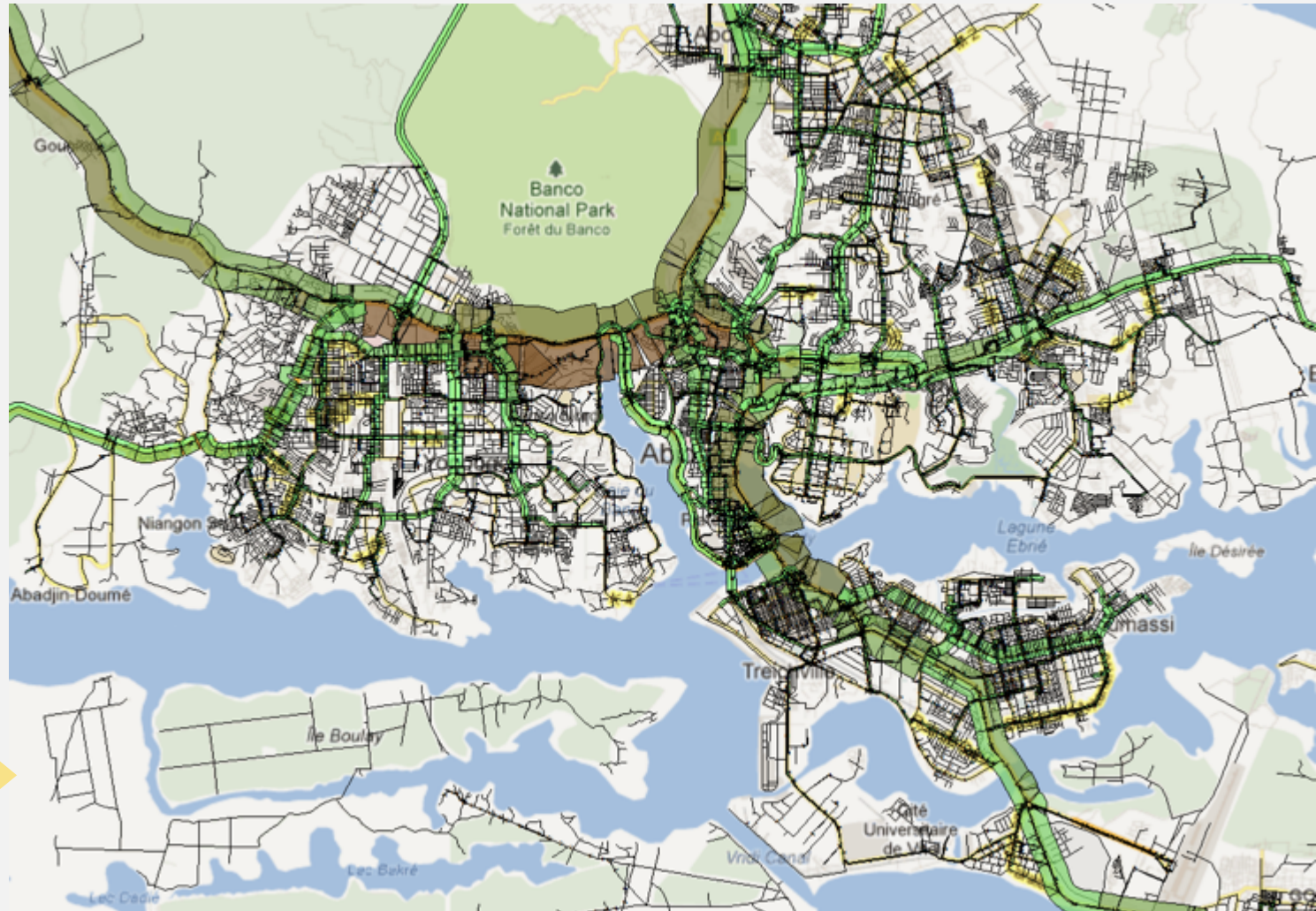
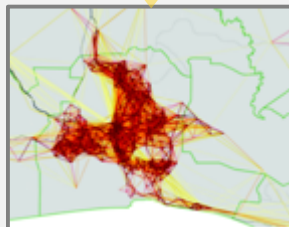
# DATA-TO-KNOWLEDGE LIFECYCLE



# CRISP-DM MODEL CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

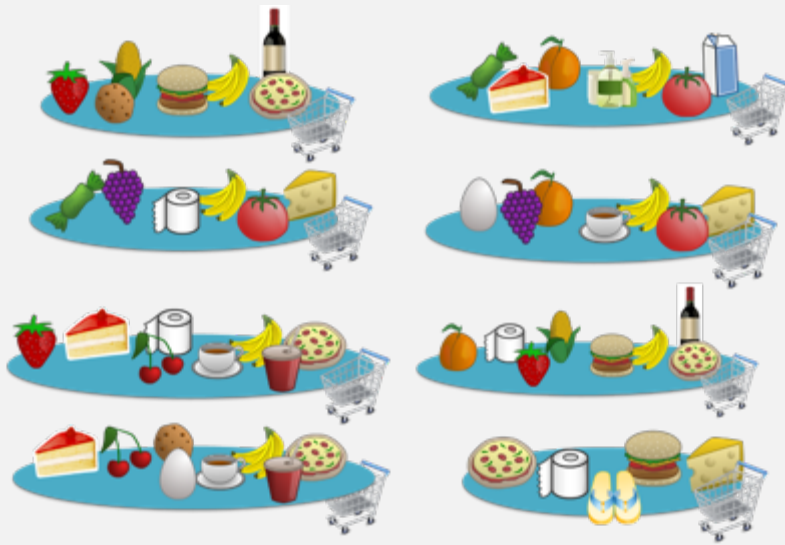


# UNDERSTANDING CITY USAGE



# ECONOMICAL TRANSACTIONS

- E.g. personal shopping data
  - Contain the history of purchases of the individual
  - Not only how much she buys, but also what and when...

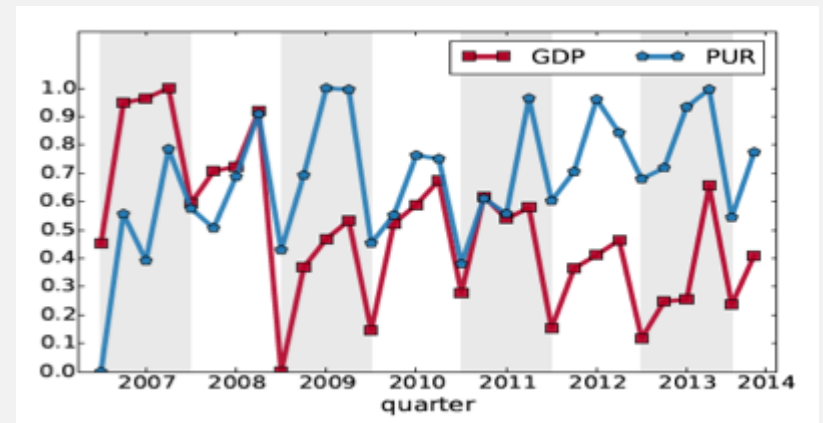
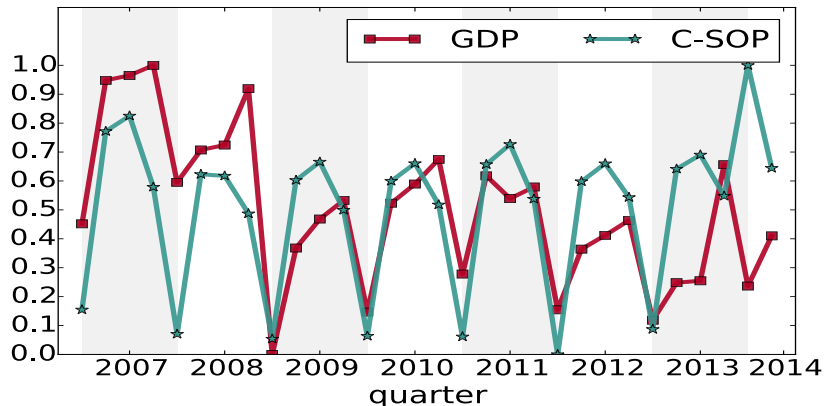
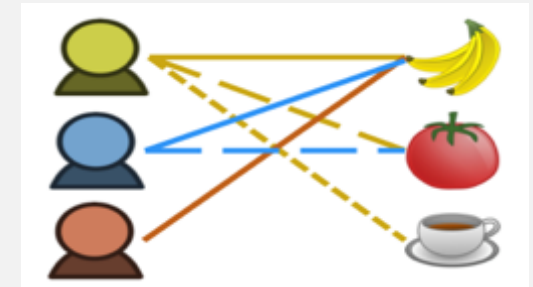




# NOWCASTING GDP & WELL-BEING

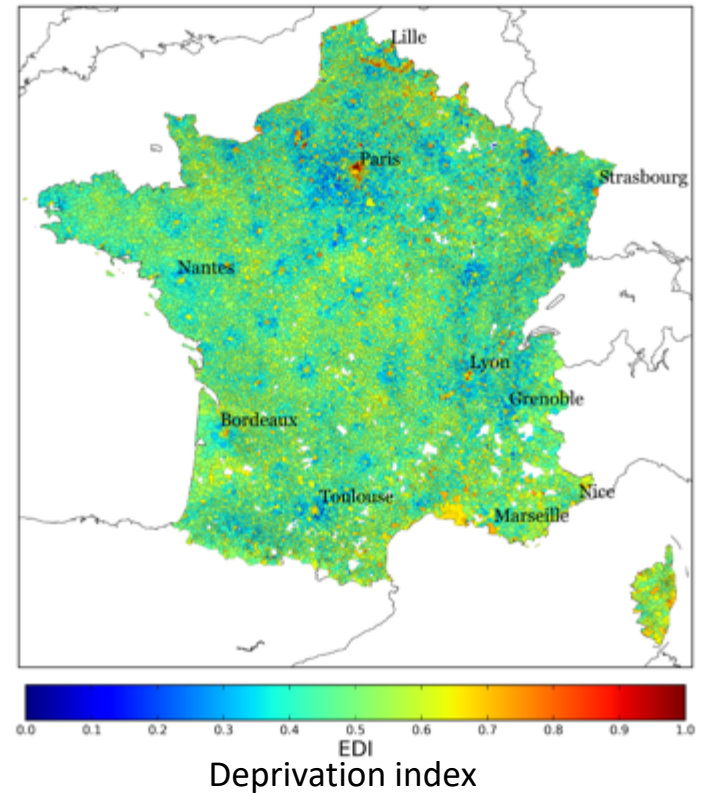
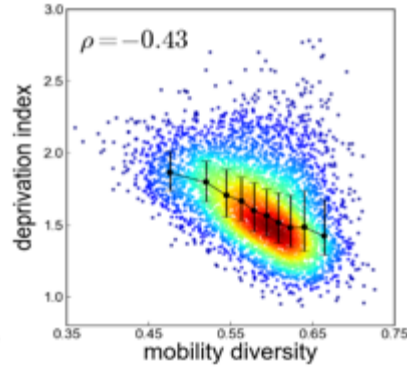
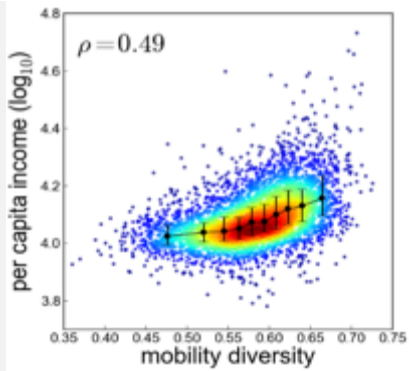
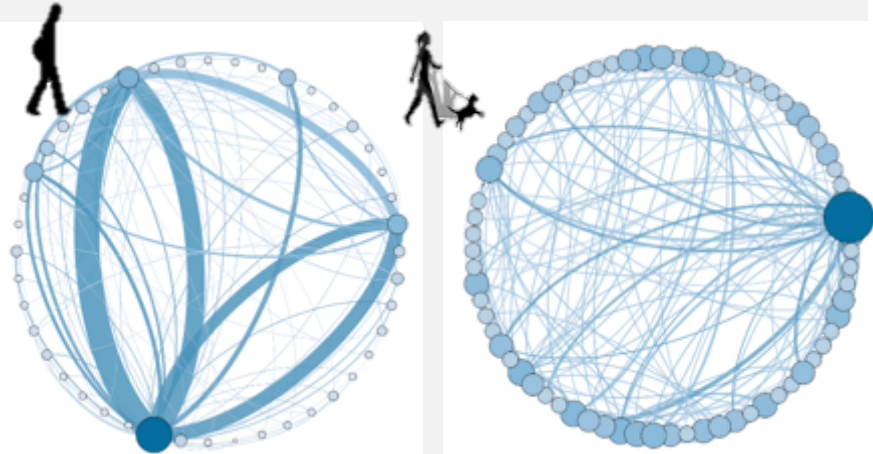
## Customers-Product and Sophistication

- Customers are sophisticated if they purchase sophisticated products.
- Products are sophisticated if they are bought by sophisticated customers.



Relation between GDP and customers sophistication (left) and product purchased (right)

# MOBILITY AS PROXY OF WELL-BEING



# REFUGEE MIGRATION IN EUROPE



**1 MILLION REFUGEES  
IN 2015** (source UNHCR)

# HUMAN GENETIC DISEASE NETWORKS

