DDAMINTRODUCTION TO BIGDATA

Docente: Patrizio Dazzi

Mail: patrizio.dazzi@isti.cnr.it

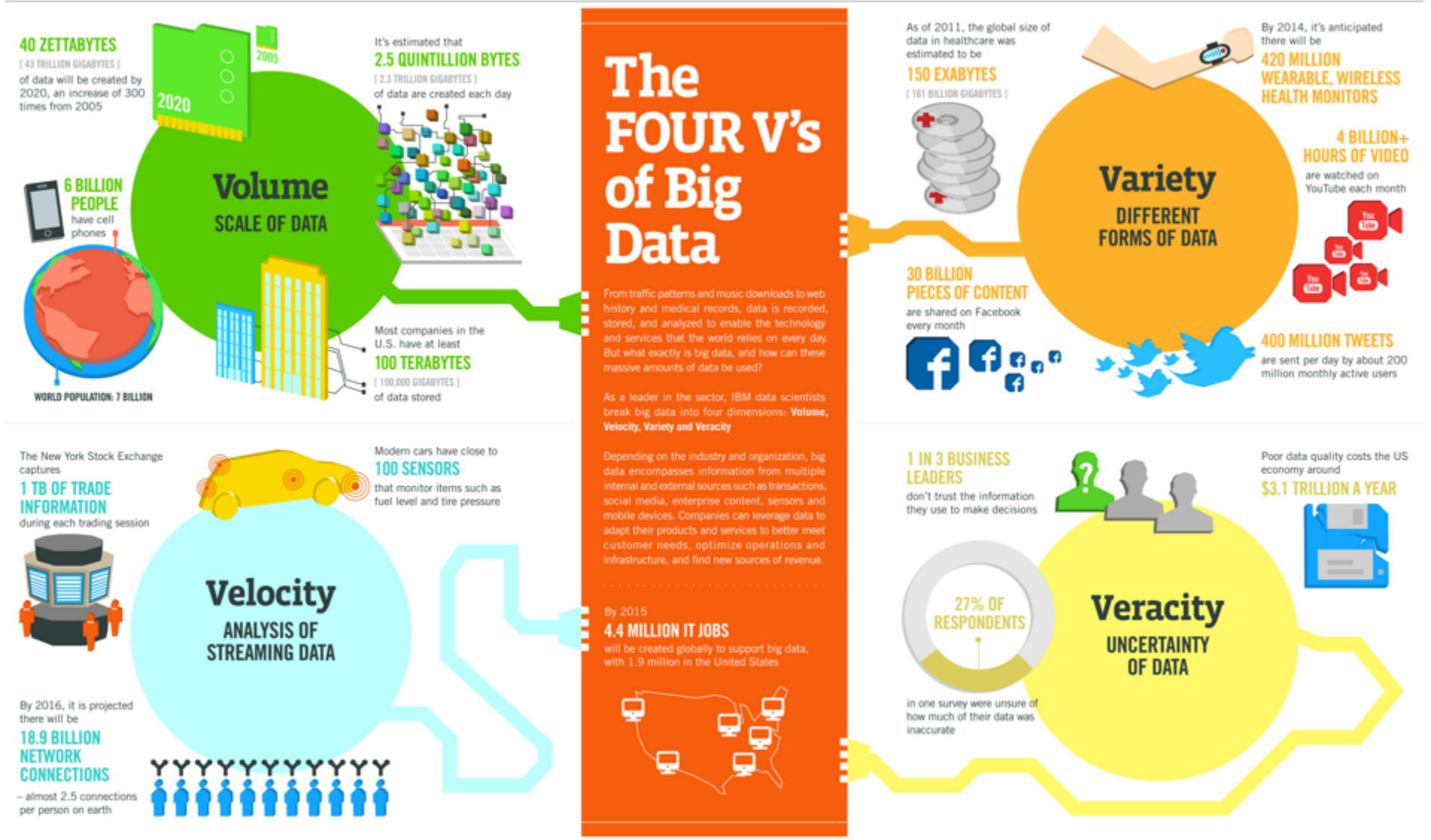

2005
Luca Bruno / AP

DIGITAL LITTLE THUMBLINGS

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


IN 60 SECONDS...
1 NEW DEFINITION IS ADDED ON urban
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
1 associatedcontent NEW ARTICLE IS PUBLISHED
THE WORLD'S LARGEST COMMUNITY CREATED CONTENT!!
6,600+ NEW PICTURES ARE UPLOADED ON flickr
50+ WordPress DOWNLOADS
695,000+ facebook STATUS UPDATES
125+ PLUGIN DOWNLOADS
1,700+ Firefox DOWNLOADS
79,364 WALL POSTS
510,040 COMMENTS
694,445 SEARCH QUERIES
168 MILLION EMAILS ARE SENT
60+ NEW BLOGS
1,500+ BLOG POSTS
70+ DOMAINS REGISTERED
600+ NEW VIDEOS
25+ HOURS TOTAL DURATION
100+ Answers.com
40+ YAHOO! ANSWERS
QUESTIONS ASKED ON THE INTERNET...
13,000+ iPhone APPLICATIONS DOWNLOADED
20,000+ NEW POSTS ON tumblr.
THE LARGEST SOCIAL READING PUBLISHING COMPANY!!
New Craigslist Ads
Google
Google Search
GO-Globe.com

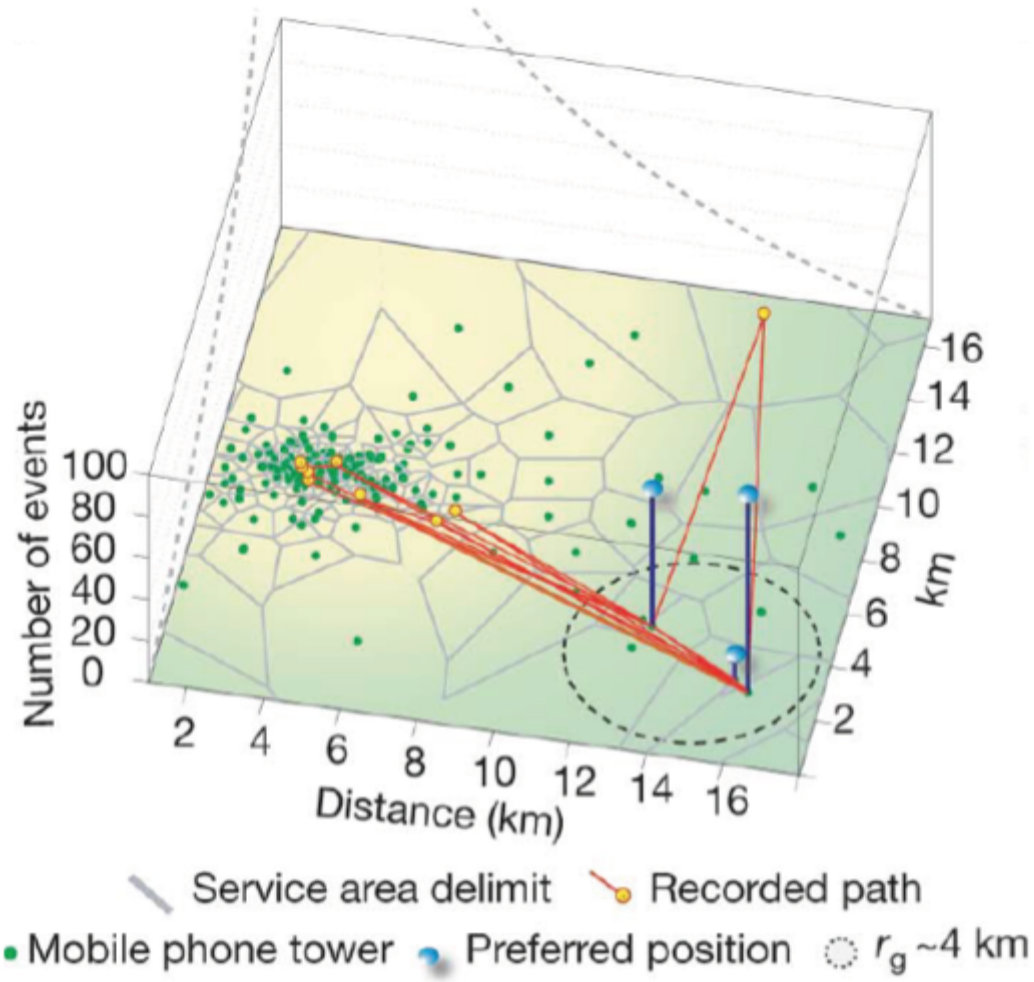
VOLUME, VELOCITY, VARIETY, VERACITY



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

IBM

CDR DATA

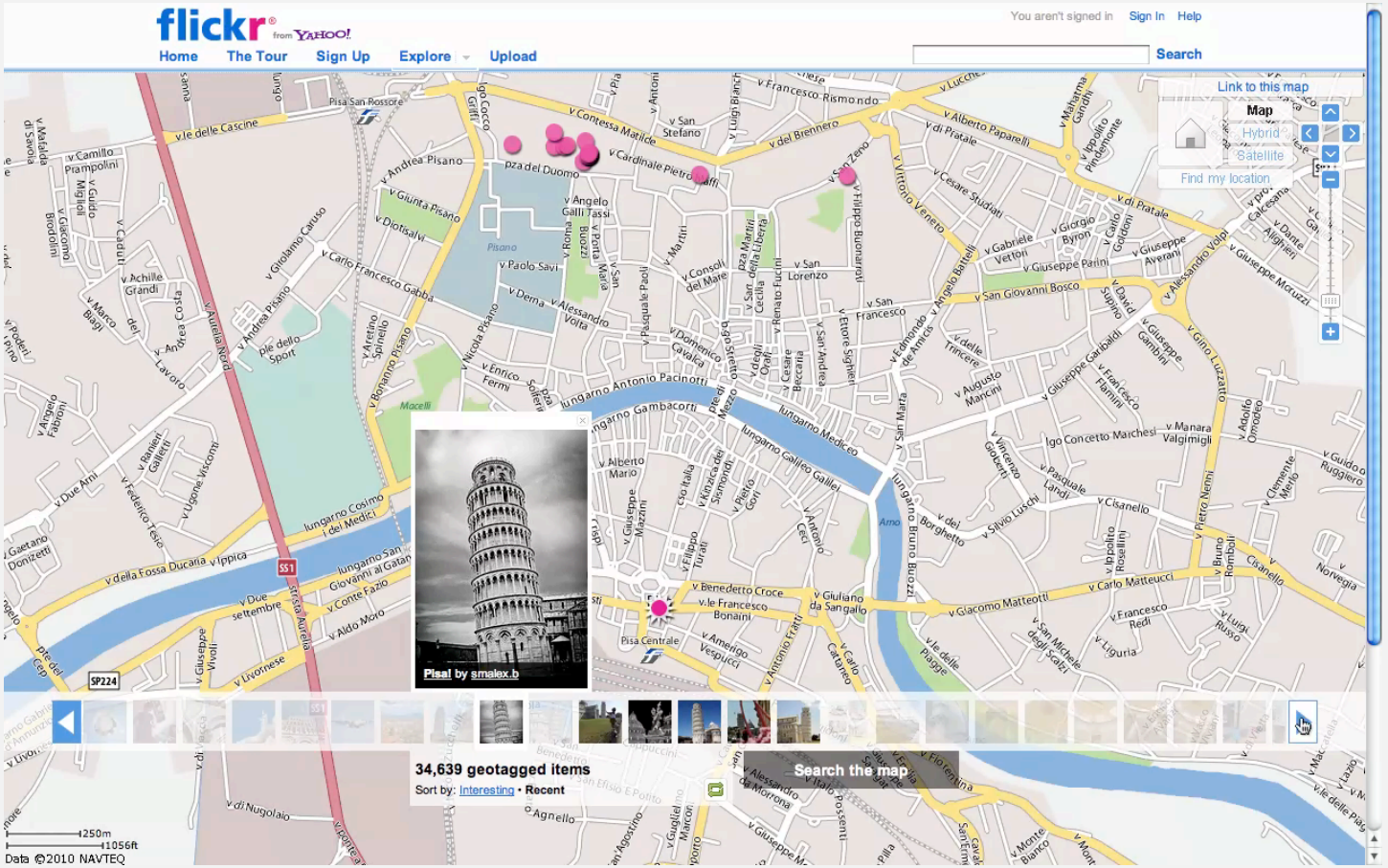


when you call

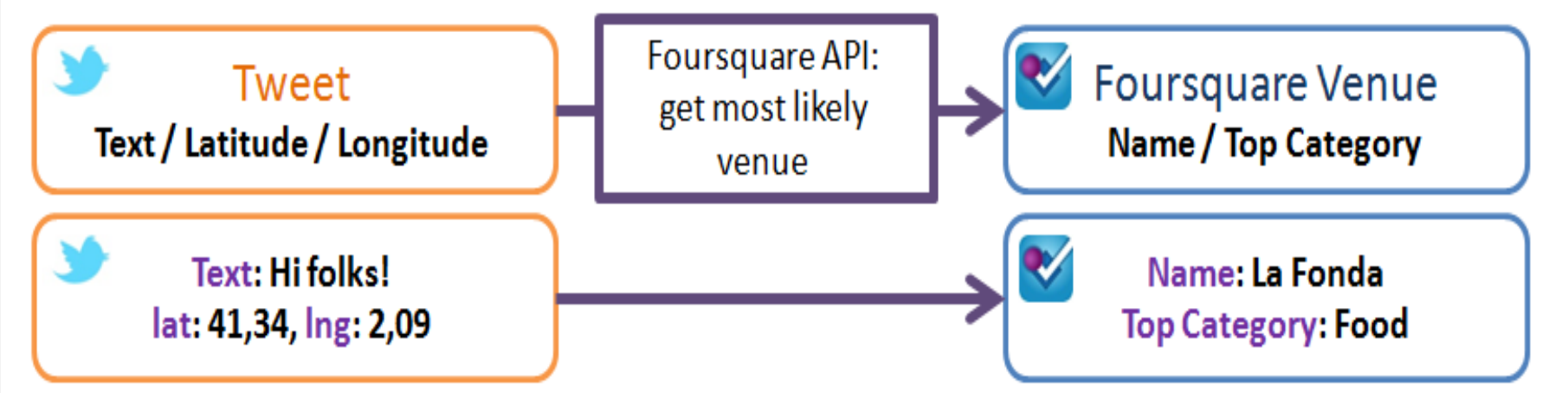
where you call

who you call

SOCIAL NETWORKS



TWITTER


Tweet
Text / Latitude / Longitude
Foursquare API: get most likely venue
Foursquare Venue
Name / Top Category
Text: Hi folks!
lat: 41,34, lng: 2,09
Name: La Fonda
Top Category: Food

GPS VEHICLE TRACKS

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

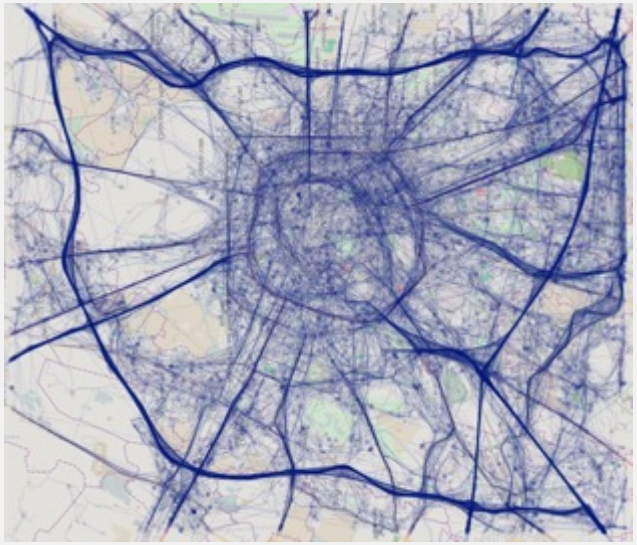
Ide;Time;Lat;Lon;Height;Course;Speed;PDOP;State;NSat

...

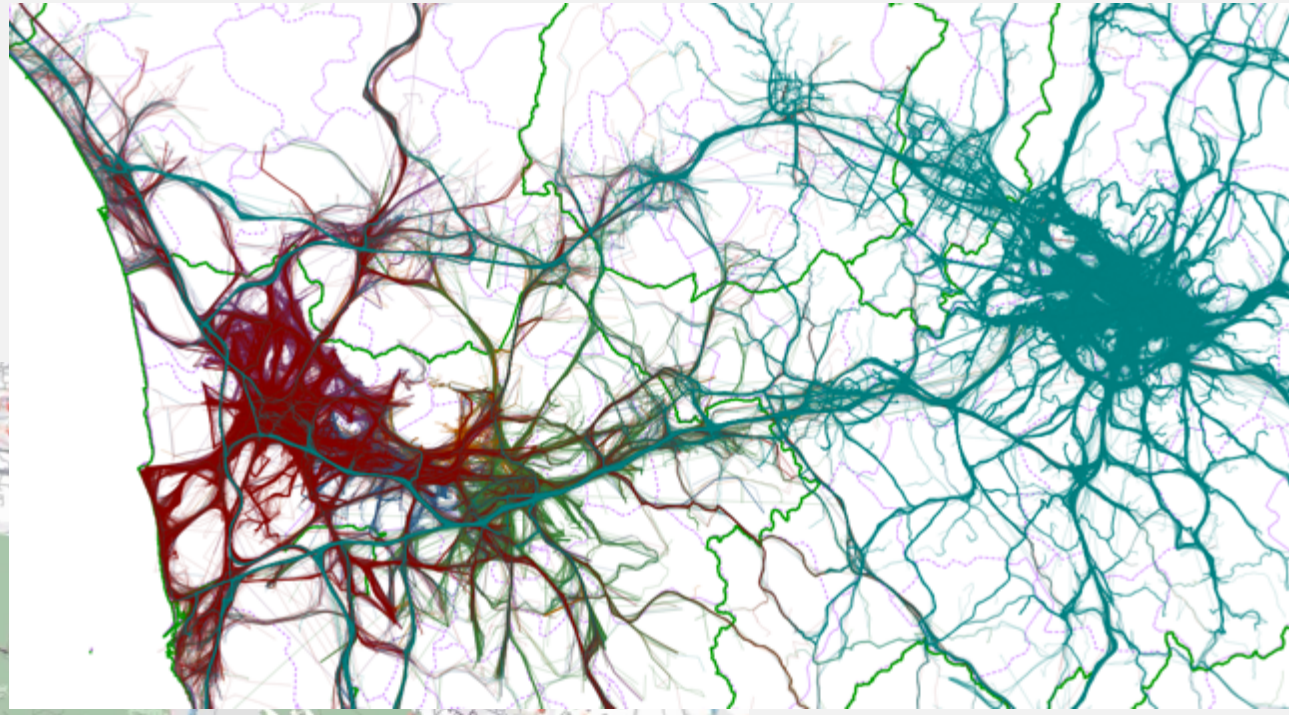
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
8;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
8;22/03/07 08:51:59;50.777415;7.205543; 68.3;112.7;25.298;3.8;1808;4
8;22/03/07 08:52:03;50.777317;7.205877; 68.8;119.8;32.447;3.8;1808;4
8;22/03/07 08:52:06;50.777185;7.206202; 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
...

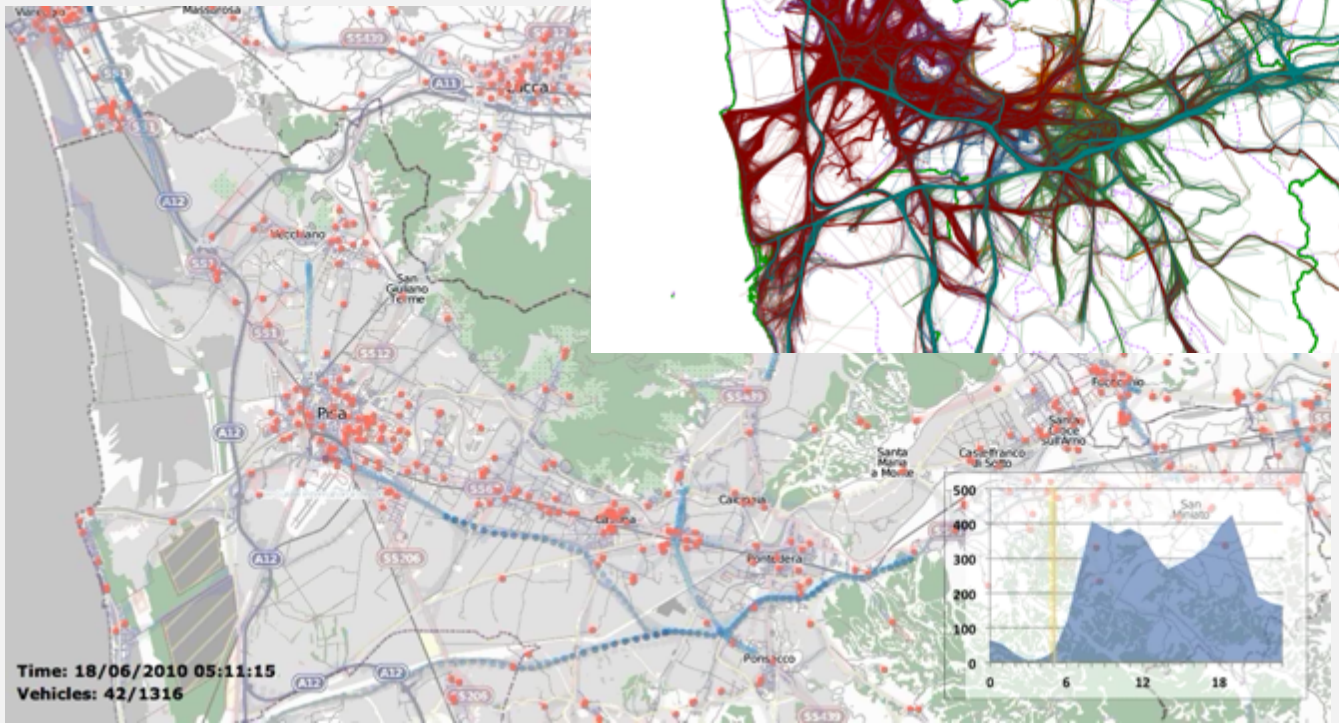
Typical sampling rate: 30-60 secs

Typical localization error: 5-10 m

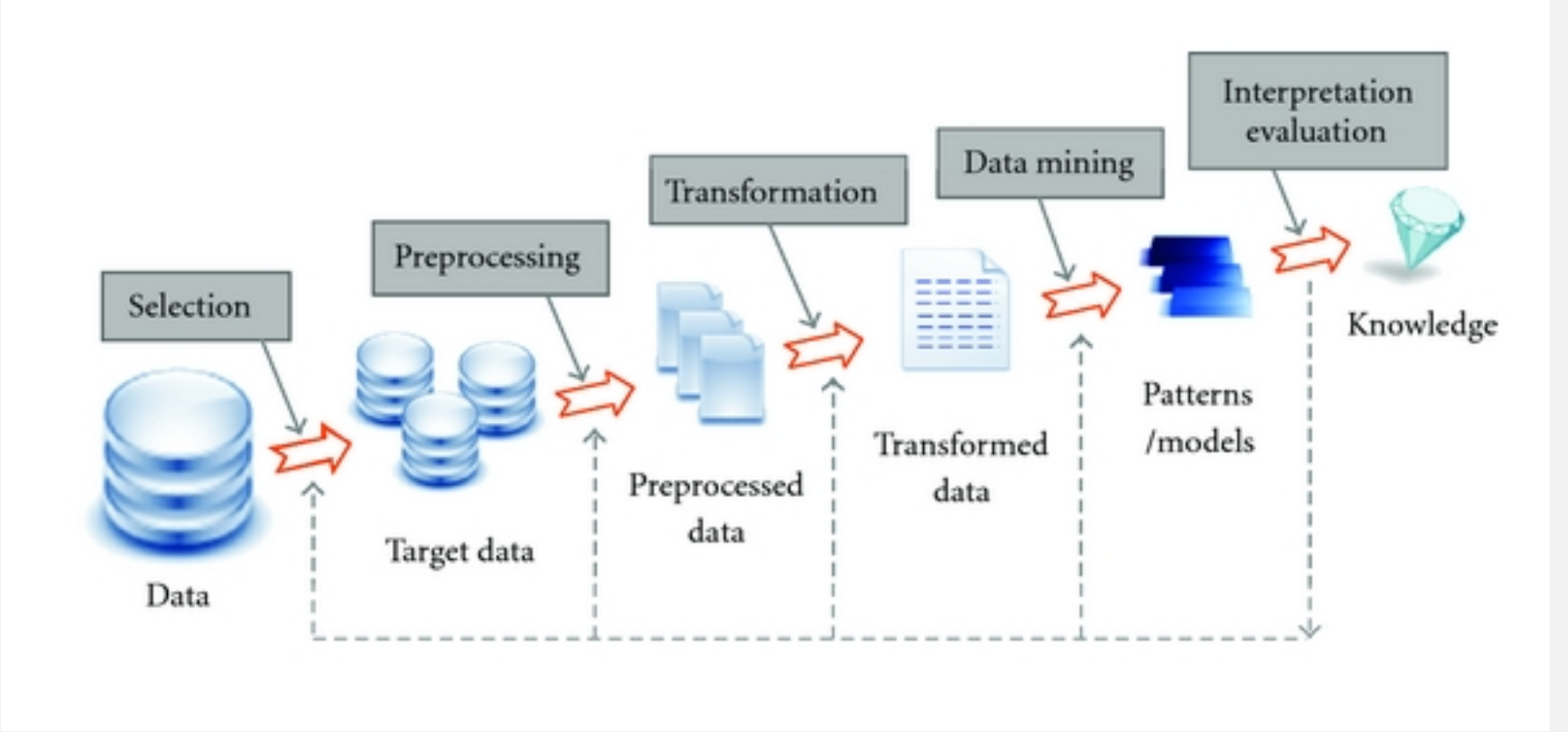


GPS vehicle tracks



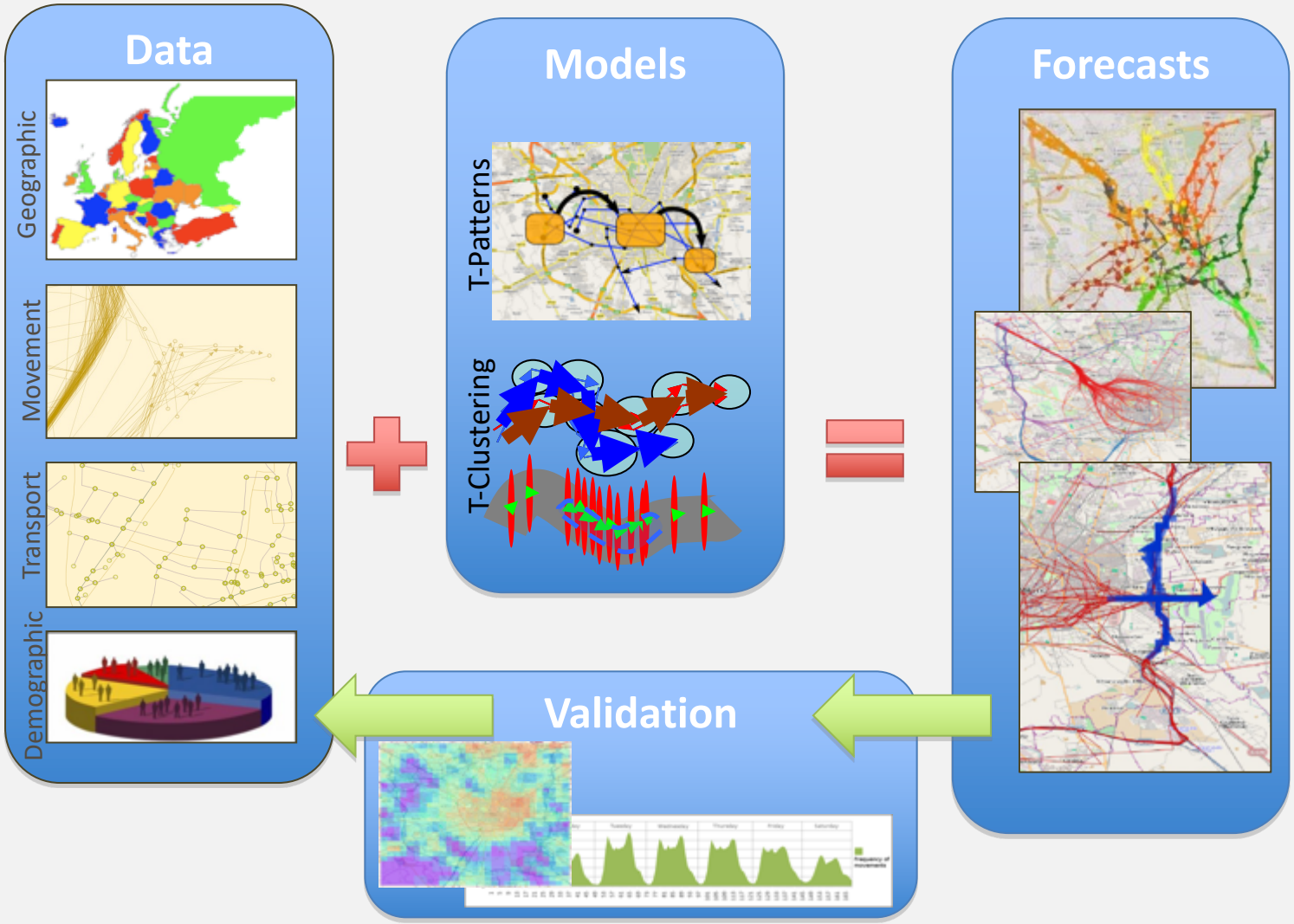


THE KDD PROCESS



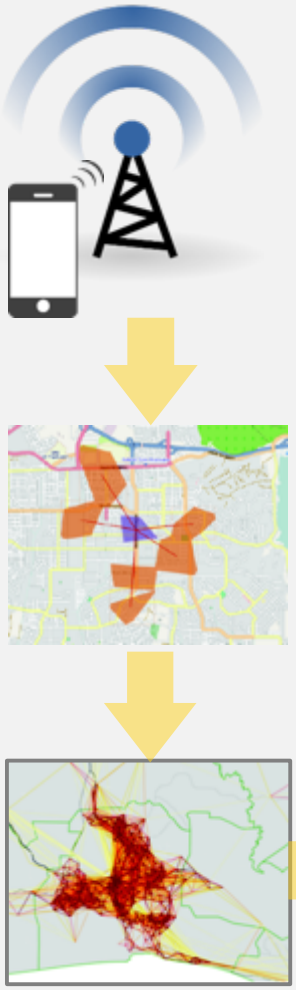
http://www.rithme.eu/img/KDprocess.png

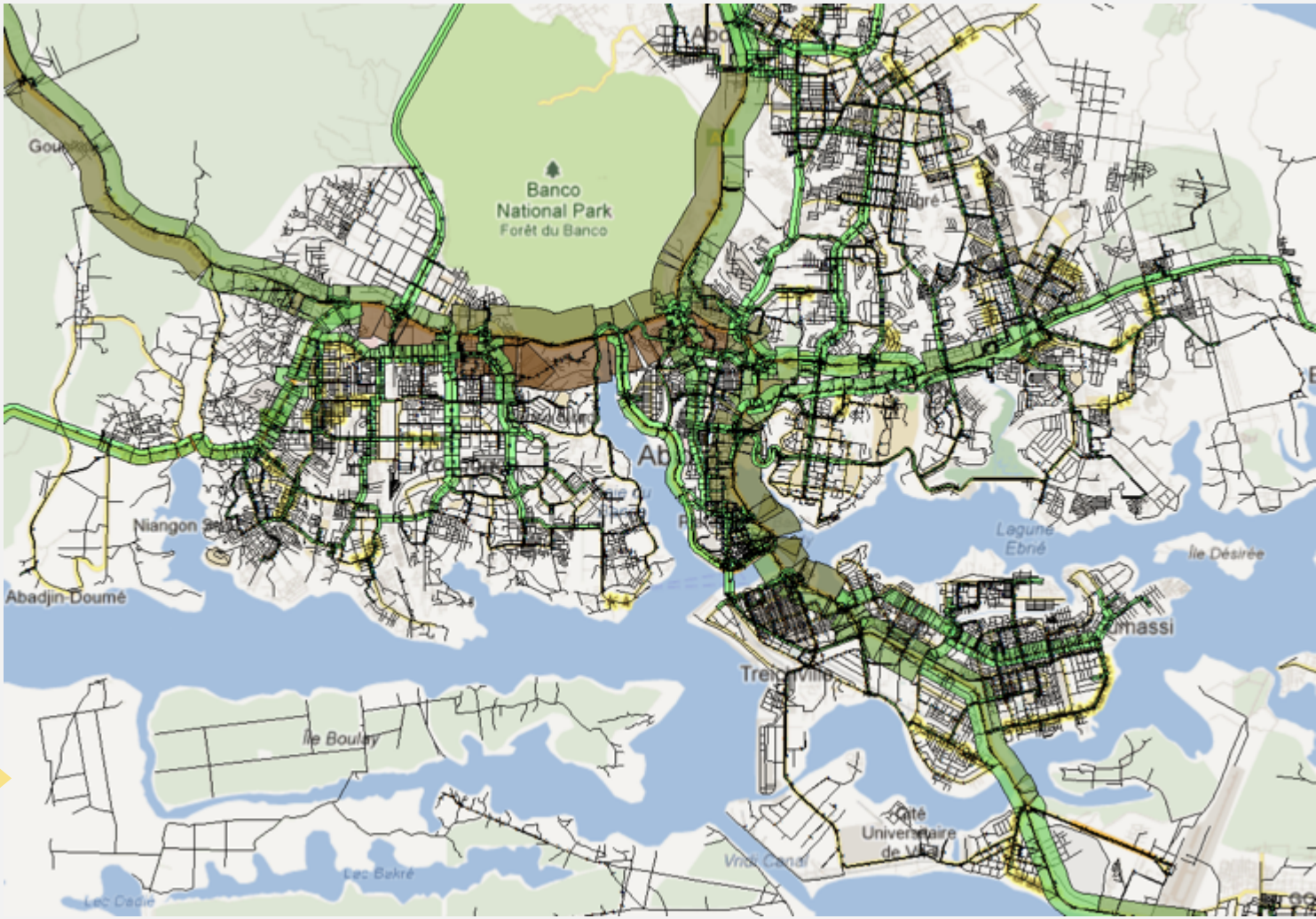
DATA-TO-KNOWLEDGE LIFECYCLE



CRISP-DM MODEL
CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING

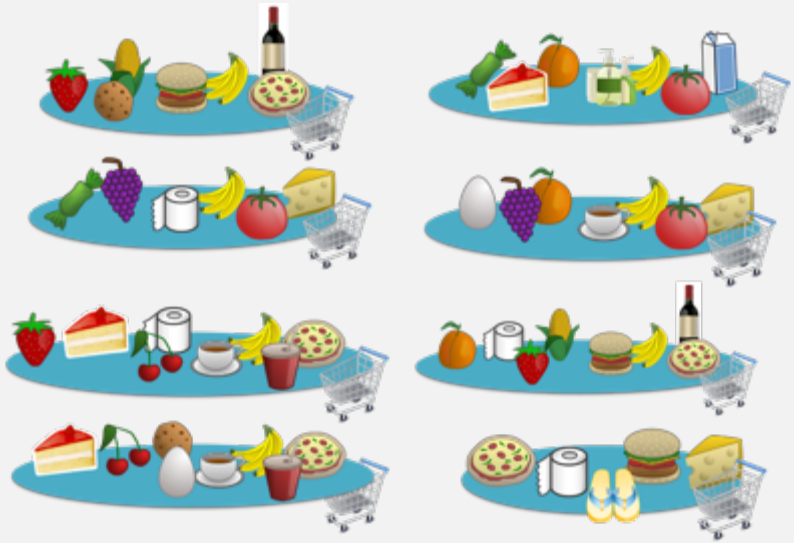
UNDERSTANDING CITY USAGE





ECONOMICAL TRANSACTIONS

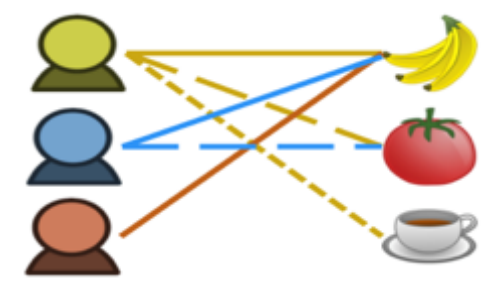
- E.g. personal shopping data
  - Contain the history of purchases of the individual
  - Not only how much she buys, but also what and when…

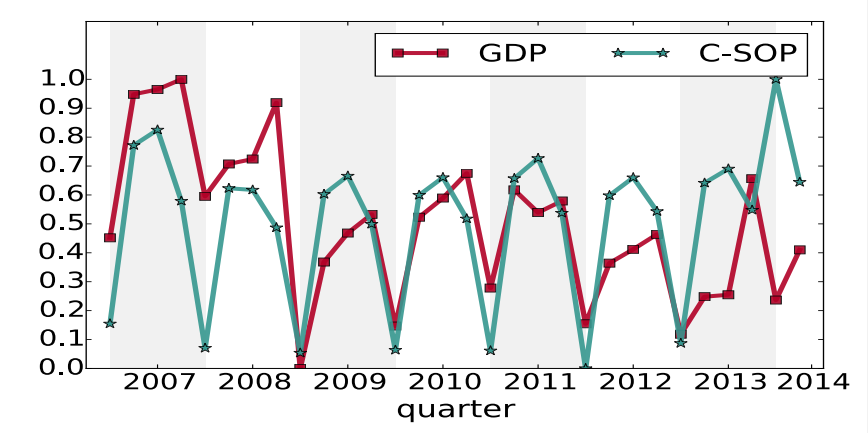


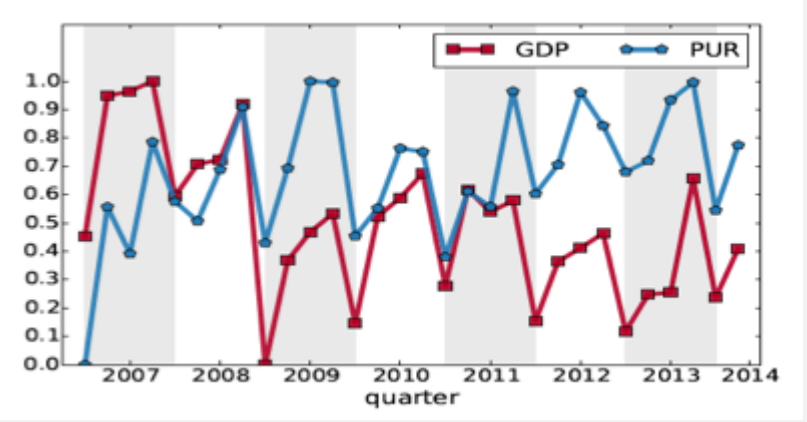
NOWCASTING GDP & WELL-BEING

Customers-Product and Sophistication

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

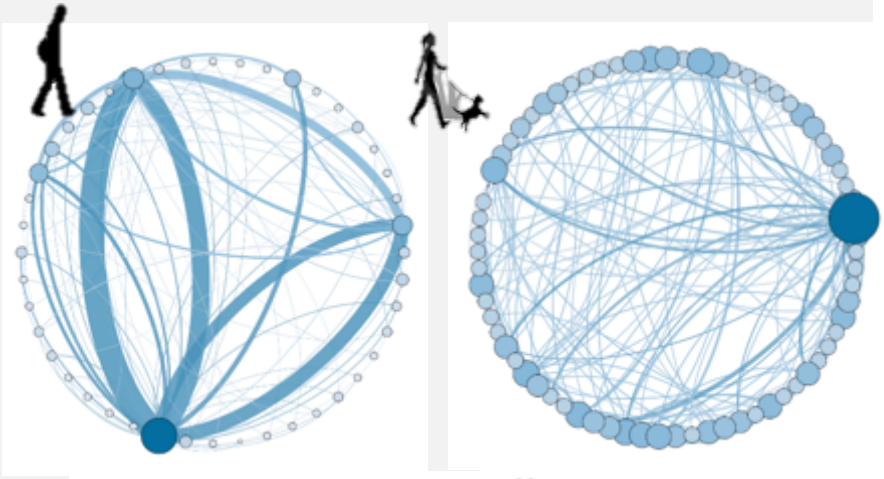


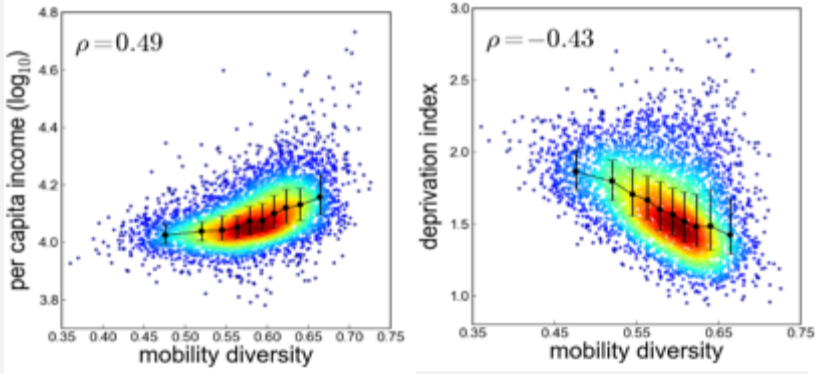


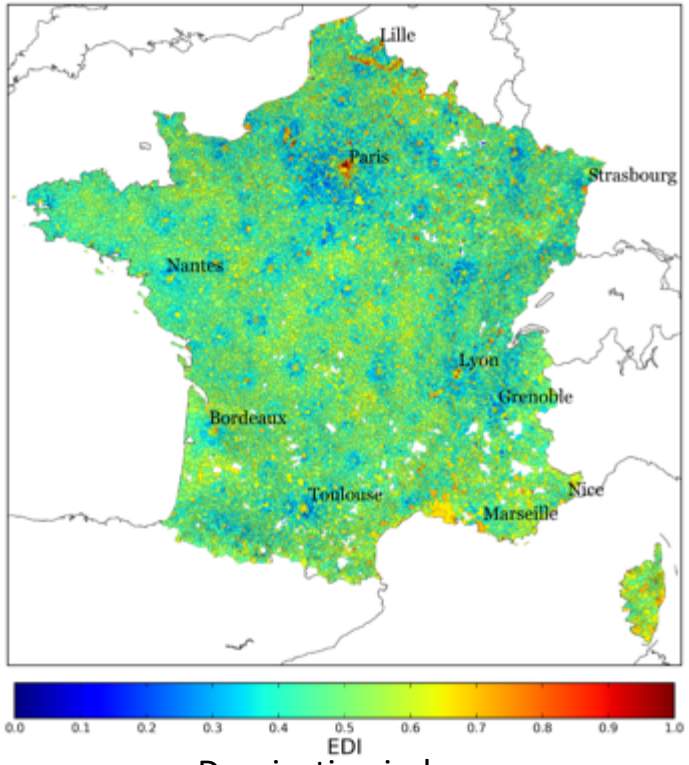


Relation between GDP and customers sophistication (left) and product purchased (right)

MOBILITY AS PROXY OF WELL-BEING







Deprivation index

REFUGEE MIGRATION IN EUROPE

1 MILLION REFUGEES IN 2015 (source UNHCR)

HUMAN GENETIC DISEASE NETWORKS