Big data Analytics & Data Mining

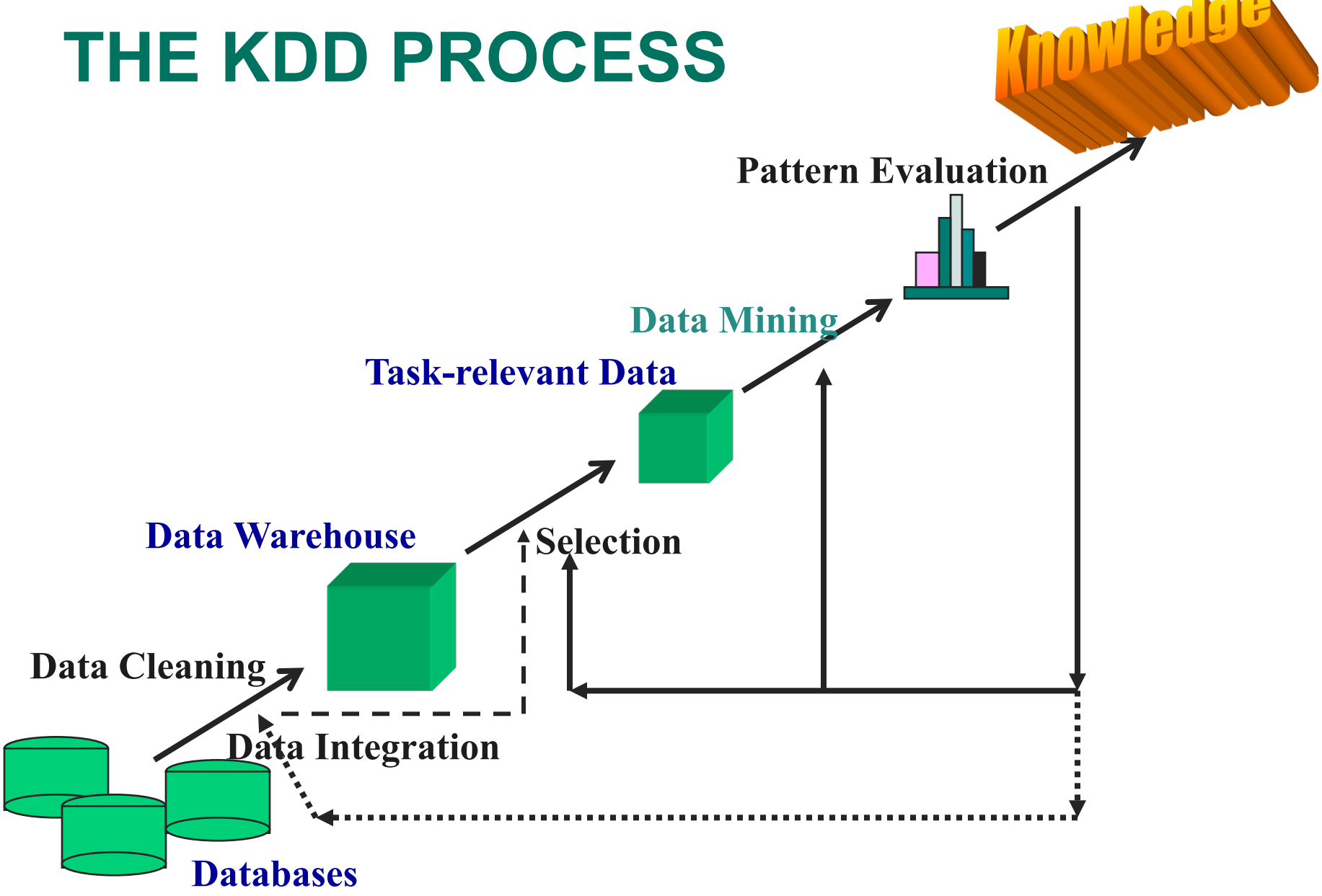
ANNA MONREALE
EMAL: ANNA.MONREALE@UNIPI.IT
DIPARTIMENTO DI INFORMATICA
UNIVERSITA' DI PISA

WHAT IS DATA MINING?

Data mining (knowledge discovery from data)

Data mining is the use of **efficient** techniques for the analysis of **very large collections of data** and the **extraction** of useful and possibly unexpected patterns in data (hidden knowledge).

THE KDD PROCESS



WHY DATA MINING

Increased Availability of Huge Amounts of Data

- point-of-sale customer data
- digitization of text, images, video, voice, etc.
- World Wide Web and Online collections

Data Too Large or Complex for Classical or Manual Analysis

- number of records in millions or billions
- high dimensional data (too many fields/features/attributes)
- often too sparse for rudimentary observations
- high rate of growth (e.g., through logging or automatic data collection)
- heterogeneous data sources

Business Necessity

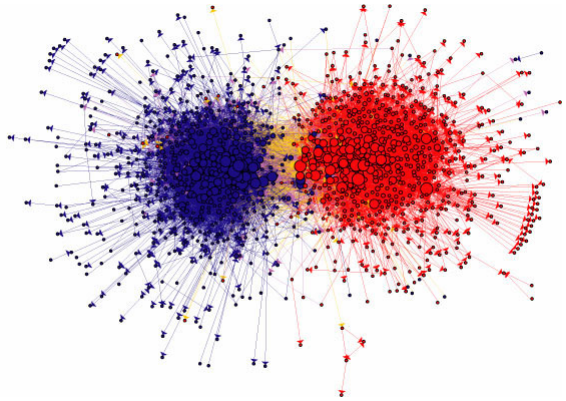
- e-commerce
- high degree of competition
- personalization, customer loyalty, market segmentation


2005
Luca Bruno / AP
2013
NBC NEWS
Michael Sohn / AP

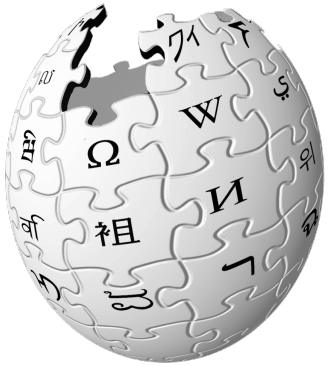
Big data proxies of social life

SHOPPING PATTERNS & LYFESTYLE

RELATIONSHIPS & SOCIAL TIES

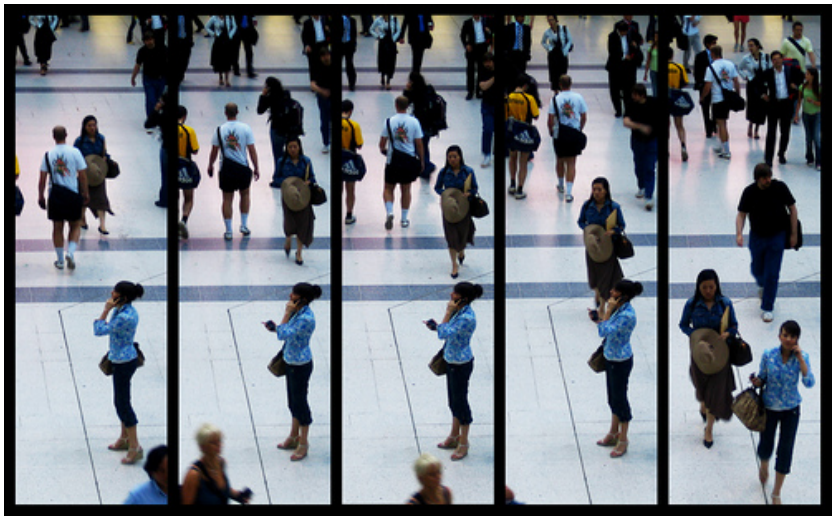


DESIRES, OPINIONS, SENTIMENTS



WIKIPEDIA
The Free Encyclopedia

MOVEMENTS



SOURCES OF DATA

Business Transactions

- widespread use of bar codes => storage of millions of transactions daily (e.g., Walmart: 2000 stores => 20M transactions per day)
- most important problem: effective use of the data in a reasonable time frame for competitive decision-making
- e-commerce data

Scientific Data

- data generated through multitude of experiments and observations
- examples, geological data, satellite imaging data, NASA earth observations
- rate of data collection far exceeds the speed by which we analyze the data

SOURCES OF DATA

Financial Data

- company information
- economic data (GNP, price indexes, etc.)
- stock markets

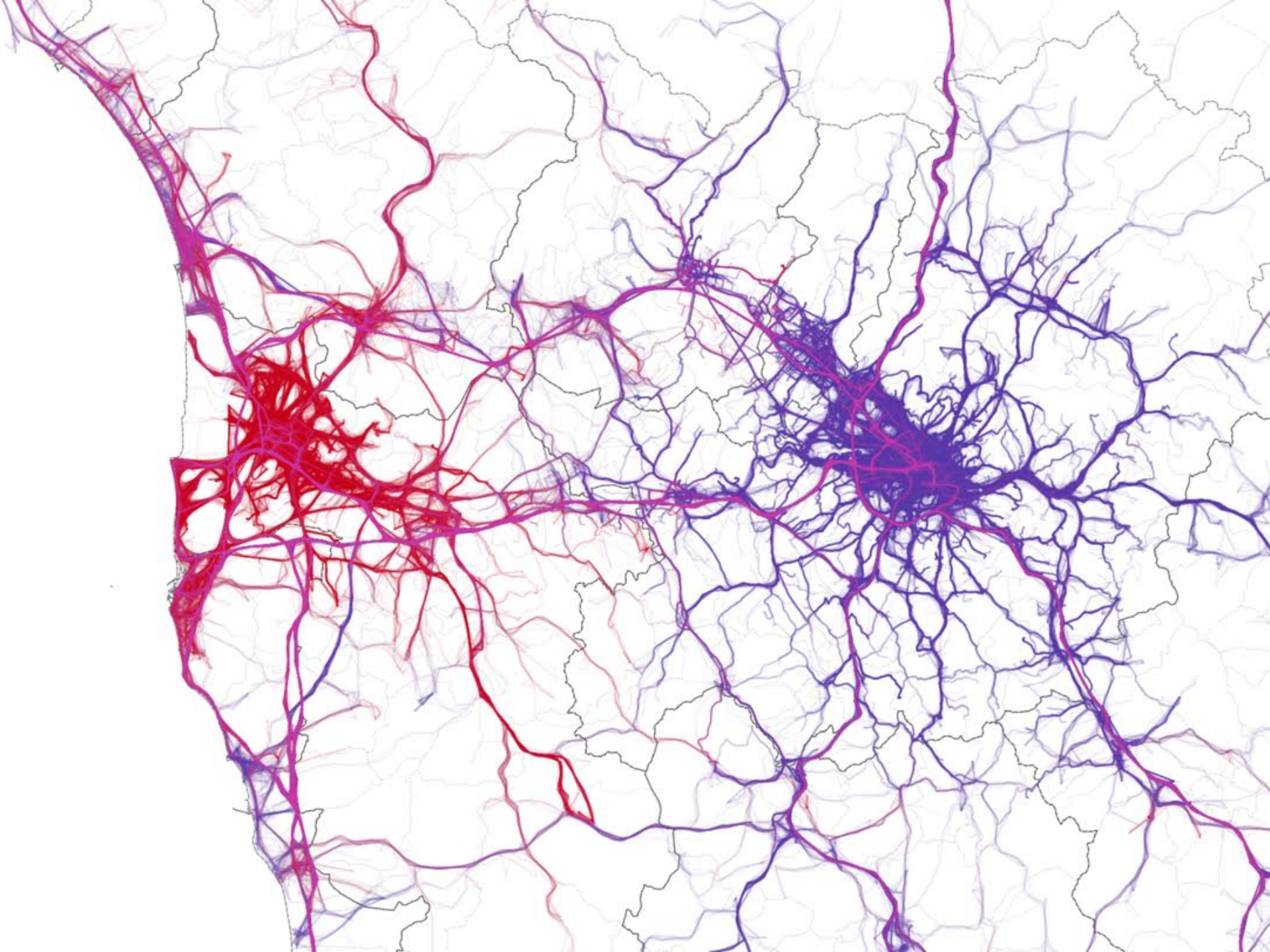
Personal / Statistical Data

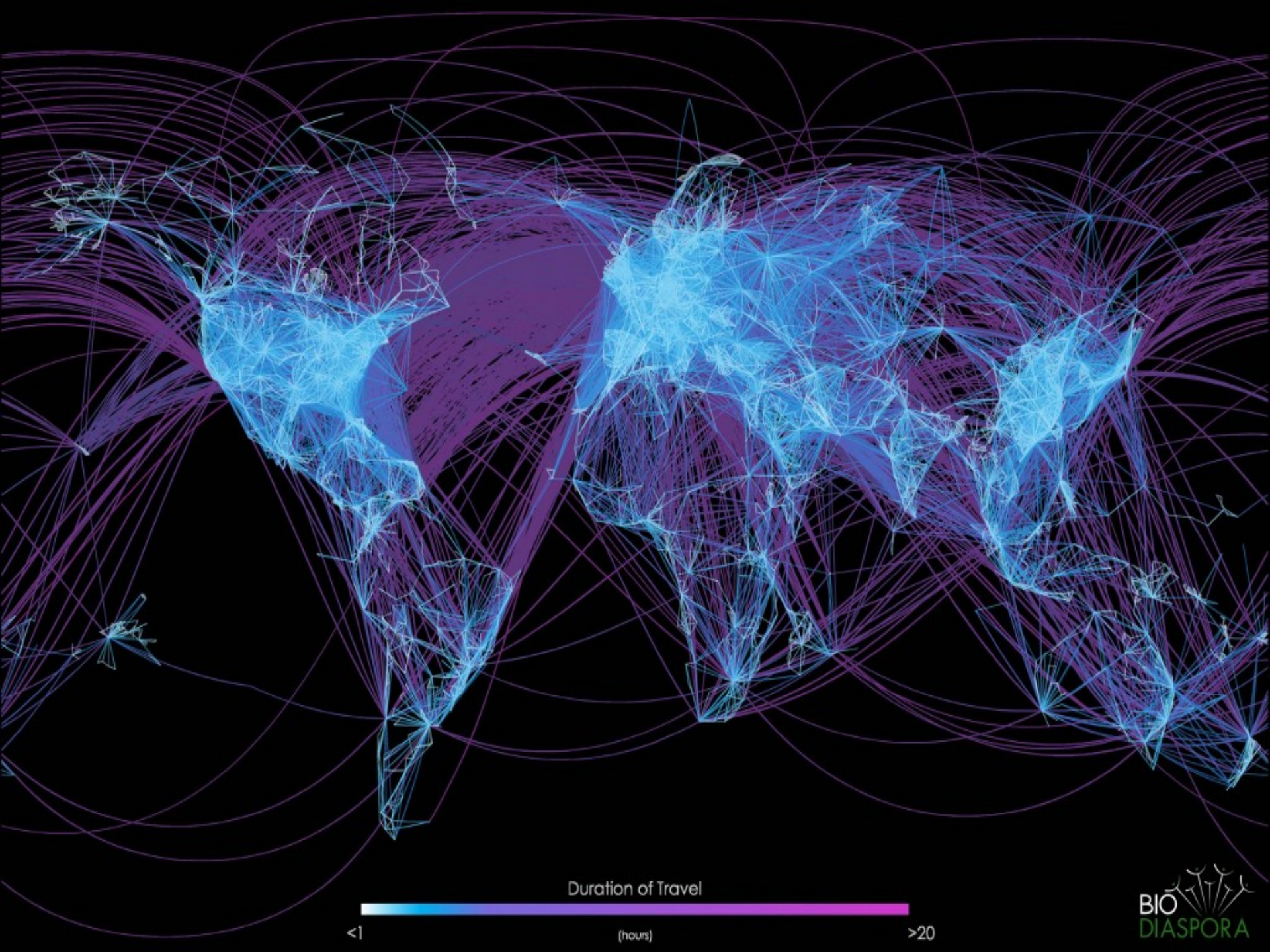
- government census
- medical histories
- customer profiles
- demographic data
- data and statistics about sports and athletes

SOURCES OF DATA

World Wide Web and Online Repositories

- email, news, messages
- Web documents, images, video, etc.
- link structure of of the hypertext from millions of Web sites
- Web usage data (from server logs, network traffic, and user registrations)
- online databases, and digital libraries
- Social network data




Duration of Travel
<1
(hours)
>20
BIO
DIASPORA

SOCIAL DATA MINING

MAKING SENSE OF BIG DATA

DATA MINING TECHNIQUES & BIG DATA ANALYTICS

Basic Techniques

- Classification
- Clustering
- Pattern Mining & Association Rules

Specific contexts

- Social Network analysis
- Mobility Data Analysis
- Social Media Analysis
- Privacy by design techniques

MARKET ANALYSIS

MARKET ANALYSIS

Where are the data sources for analysis?

- Credit card transactions, loyalty cards, discount coupons, customer complaint calls, plus (public) lifestyle studies.

Target marketing

- Find clusters of “model” customers who share the same characteristics: interest, income level, spending habits, etc.

Determine customer purchasing patterns over time

- Conversion of single to a joint bank account: marriage, etc.

Cross-market analysis

- Associations/co-relations between product sales
- Prediction based on the association information.

MARKET ANALYSIS

Customer profiling

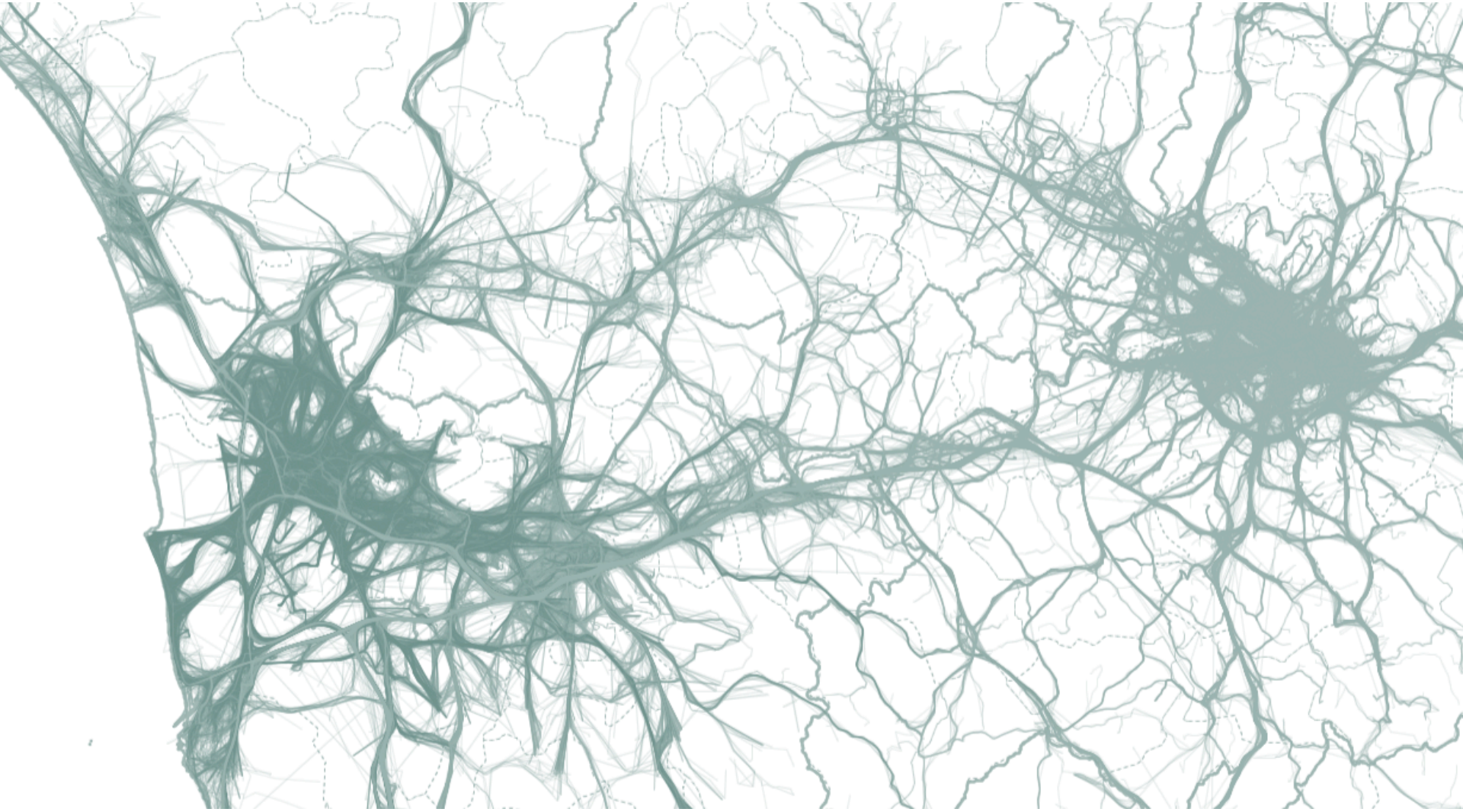
- data mining can tell you what types of customers buy what products (clustering or classification).

Identifying customer requirements

- identifying the best products for different customers
- use prediction to find what factors will attract new customers

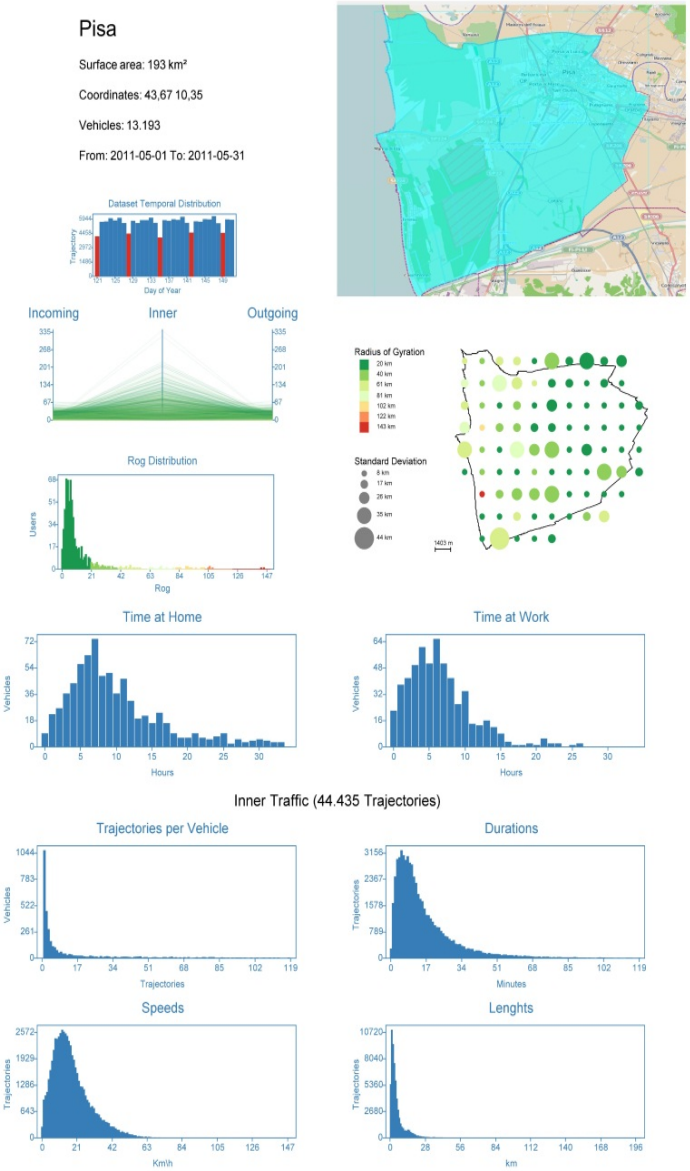
Summary information

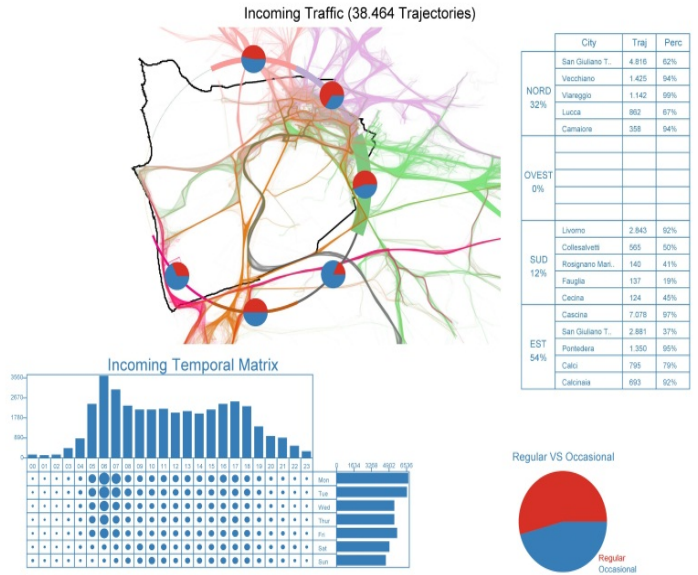
- various multidimensional summary reports;
- statistical summary information (data central tendency and variation)



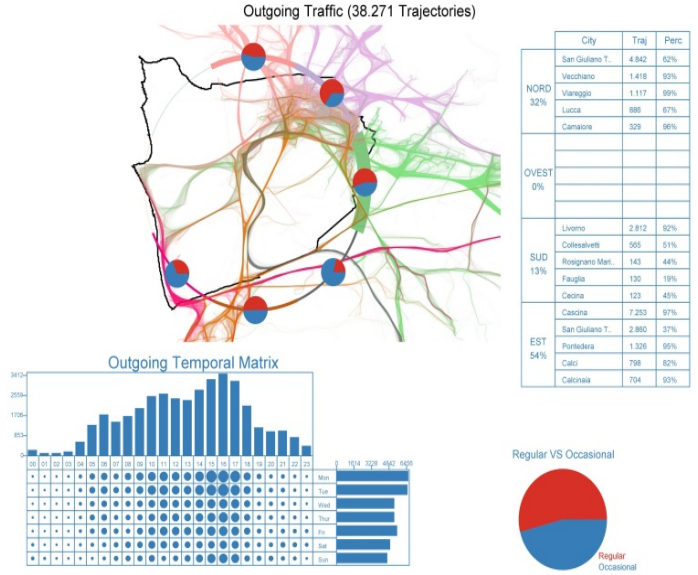
MOBILITY ATLAS OF CITIES

MOBILITY ATLAS OF MANY CITIES



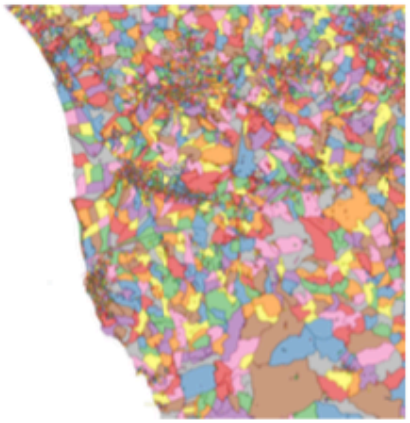
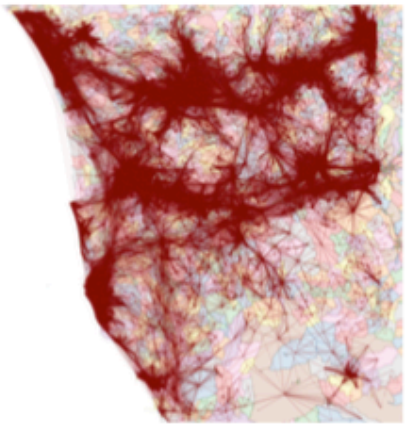
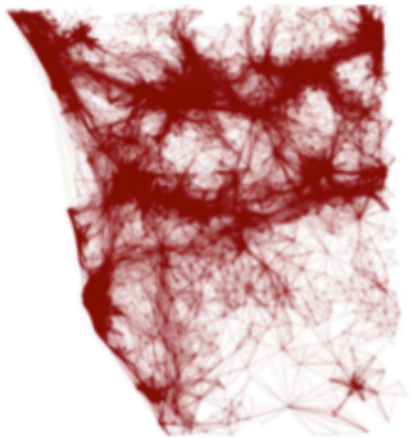
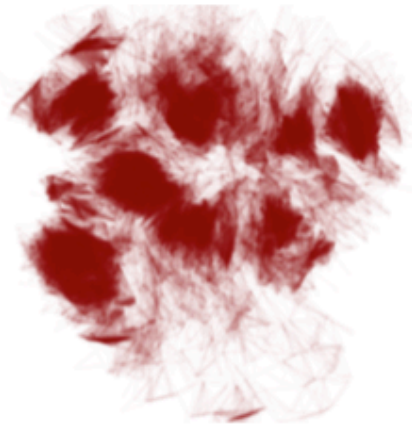
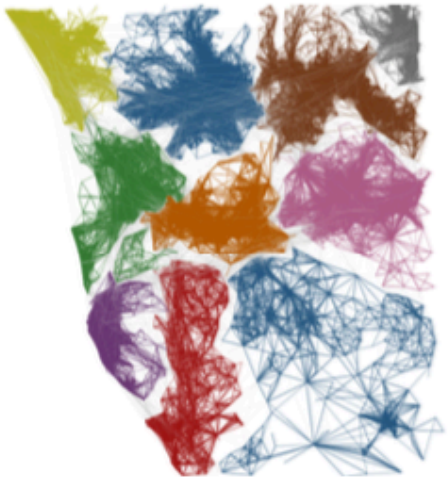


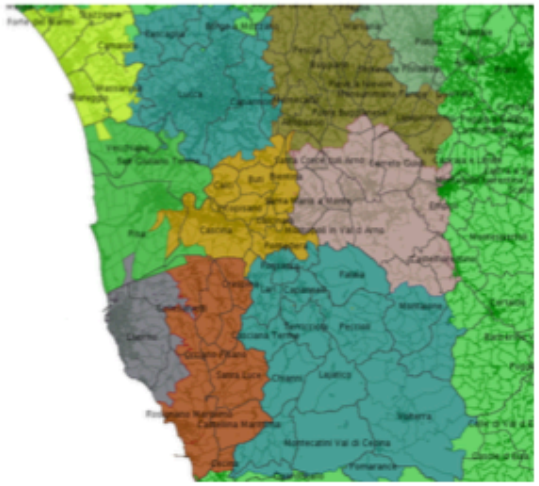
	City	Traj	Perc
NORD 32%	San Giuliano T..	4.816	62%
	Vecchiano	1.425	94%
	Viareggio	1.142	99%
	Lucca	862	67%
	Camaiore	358	94%
OVEST 0%			
SUD 12%	Livorno	2.843	92%
	Collesalvetti	565	50%
	Rosignano Mari..	140	41%
	Fauglia	137	19%
	Cecina	124	45%
EST 54%	Cascina	7.078	97%
	San Giuliano T..	2.881	37%
	Pontedera	1.350	99%
	Calci	795	79%
	Calcinaia	693	92%



	City	Traj	Perc
NORD 32%	San Giuliano T..	4.842	62%
	Vecchiano	1.418	93%
	Viareggio	1.117	99%
	Lucca	886	67%
	Camaiore	329	96%
OVEST 0%			
SUD 13%	Livorno	2.812	92%
	Collesalvetti	565	51%
	Rosignano Mari..	143	44%
	Fauglia	130	19%
	Cecina	123	45%
EST 54%	Cascina	7.253	97%
	San Giuliano T..	2.860	37%
	Pontedera	1.326	99%
	Calci	798	82%
	Calcinaia	704	93%

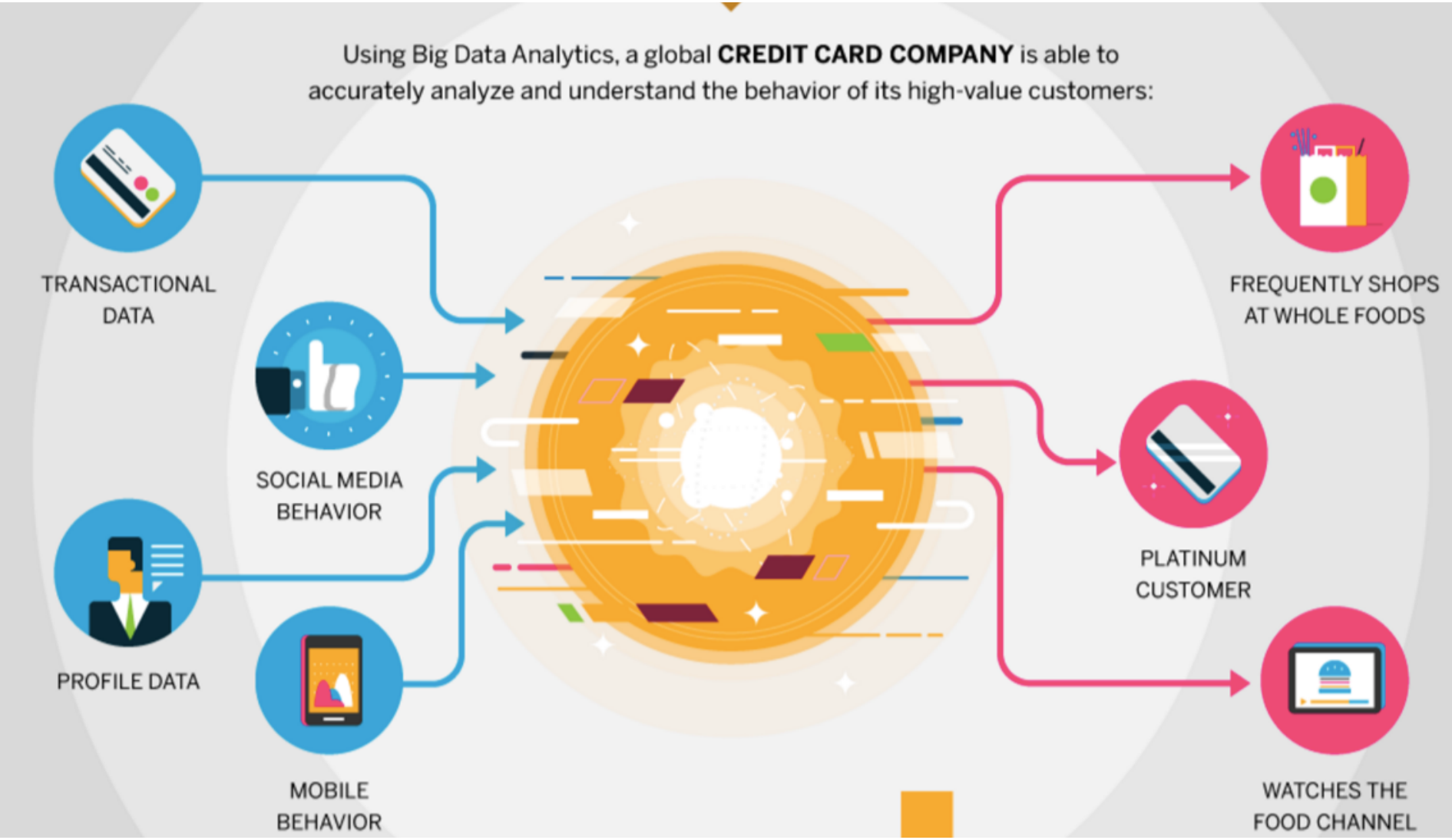
Discovering the borders of mobility

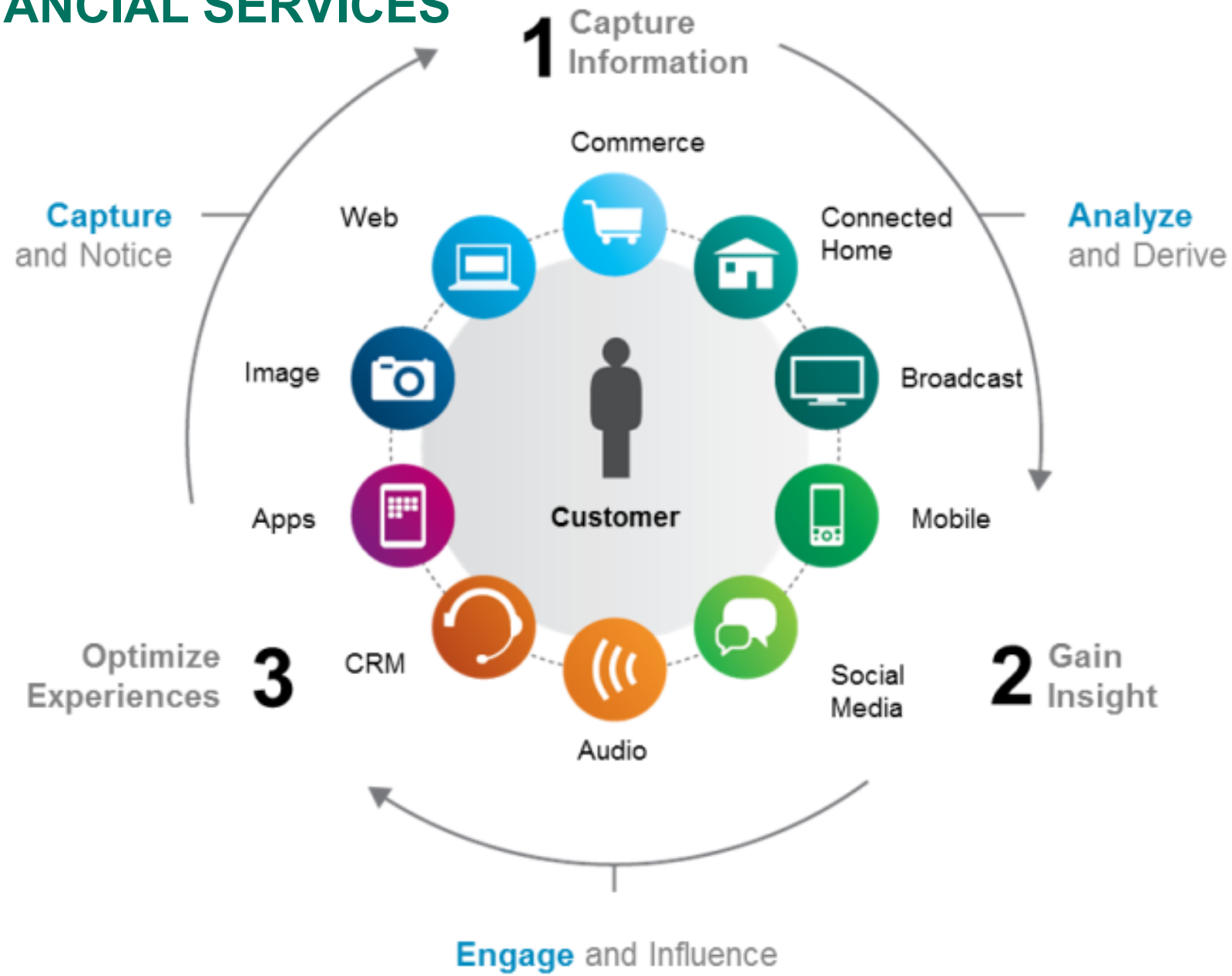


PERSONALIZED FINANCIAL SERVICES

BIG DATA: FROM CREDIT CARD TO CUSTOMER HABITS

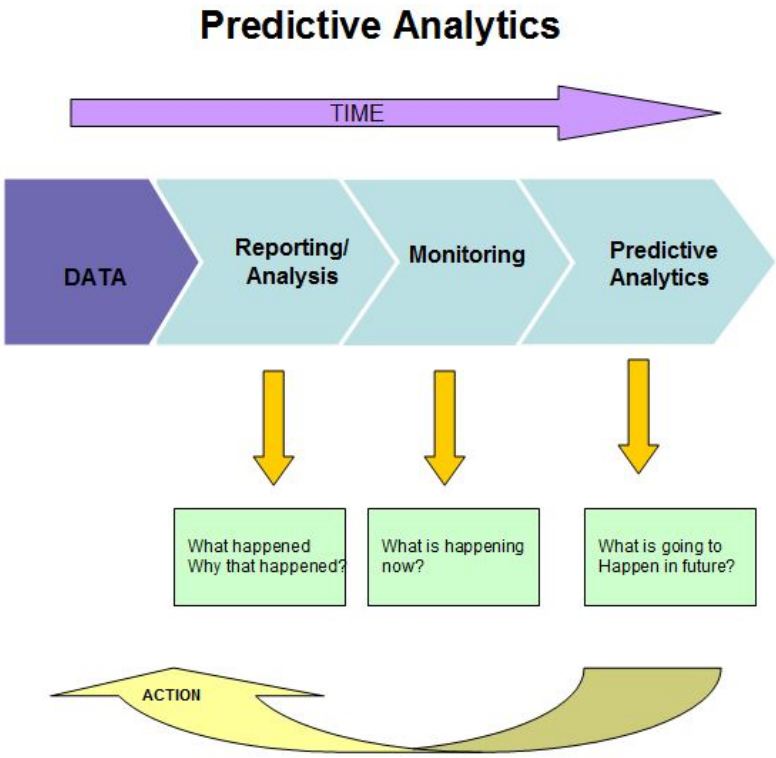


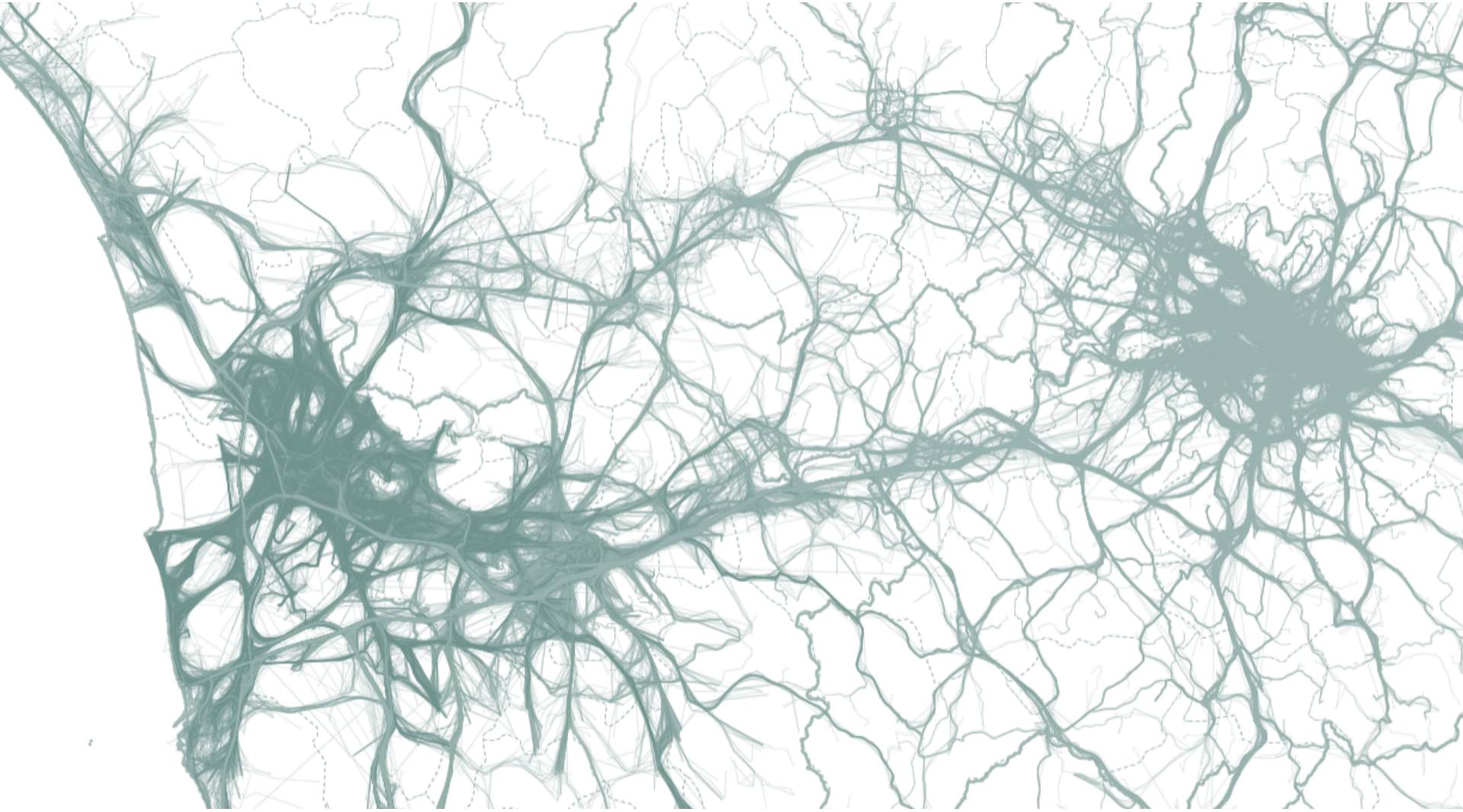
BIG DATA: NEW, MORE CAREFULLY TARGETED FINANCIAL SERVICES



DETECTION OF FUTURE FRAUD

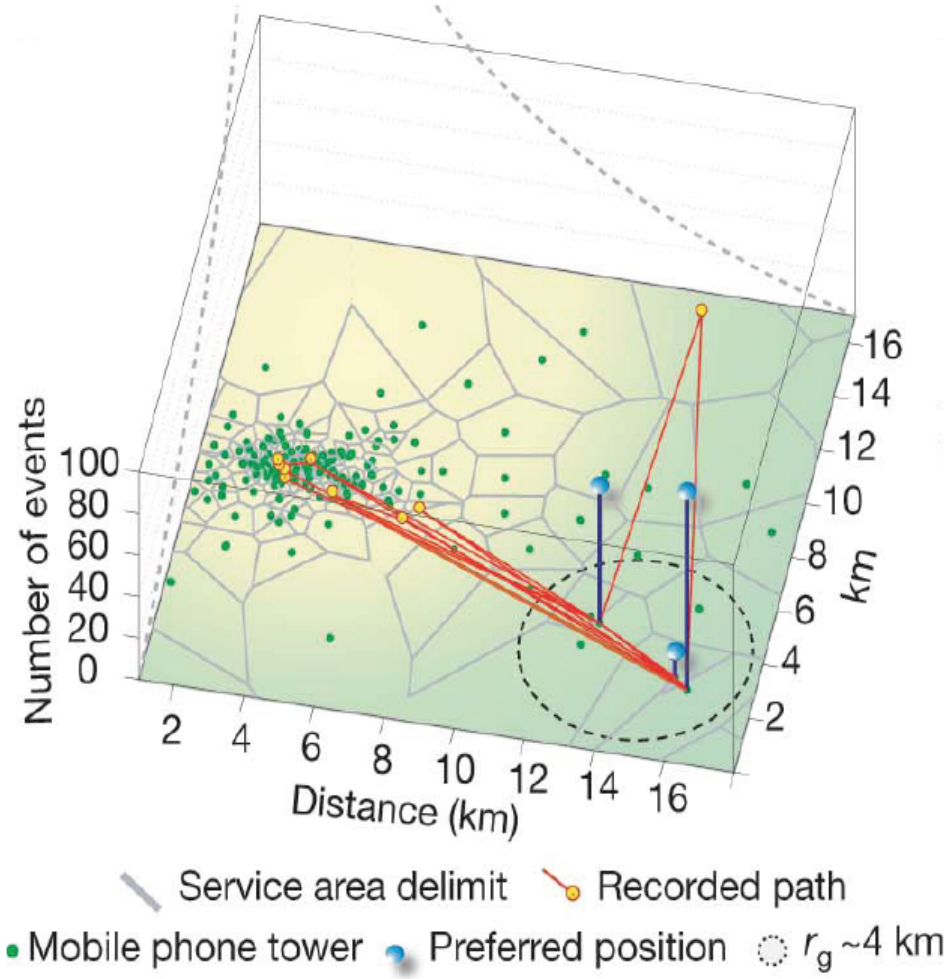
- Big Data technologies, banks can manage and analyze terabytes of historical and third-party data
- Big data capabilities for analyzing streaming data in real time.
- Create highly accurate predictive models for recognizing and preventing future fraud.
- Using this technology, banks can analyze transactions as they occur, detect fraud as it is happening and stop it before it causes serious damage.





REAL TIME DEMOGRAPHY

MOBILE PHONE (CDR) DATA

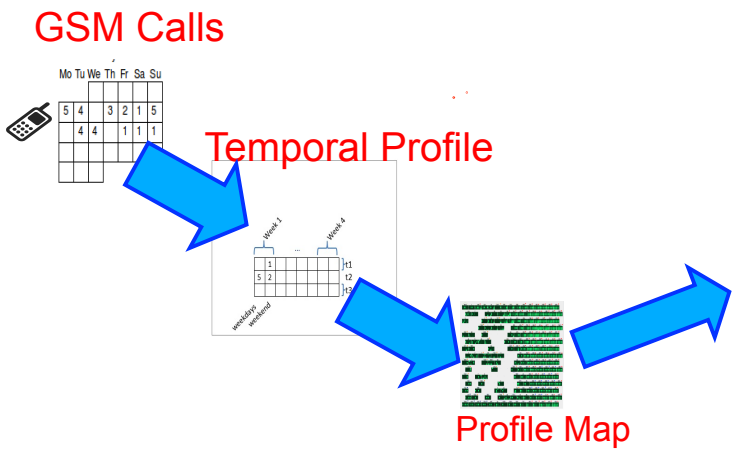


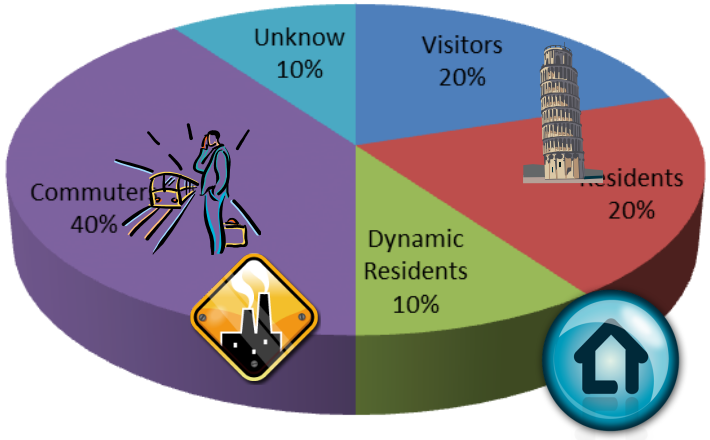
when you call

where you call

who you call

REAL TIME DEMOGRAPHICS BY MOBILE PHONE DATA





ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

KDD Lab

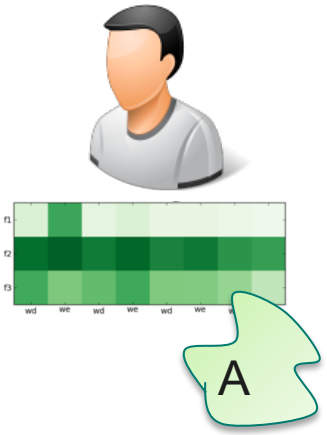
UNIVERSITÀ DI PISA

Istat

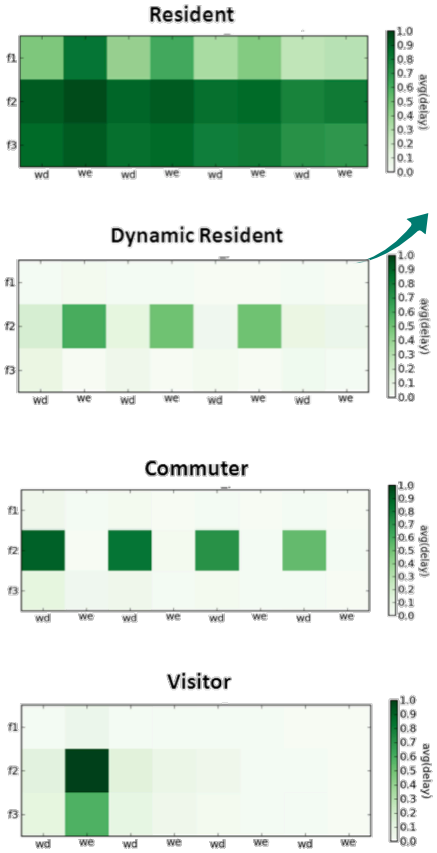
WIND
Più vicini.

CALLING PROFILES

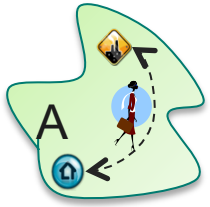
Users' Call Profiles



Classification

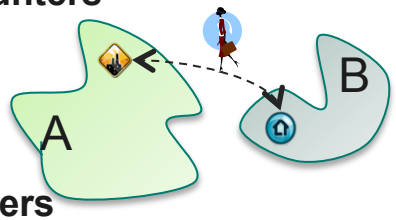


Residents

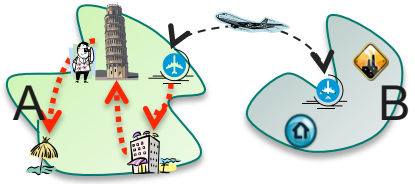


Out communters

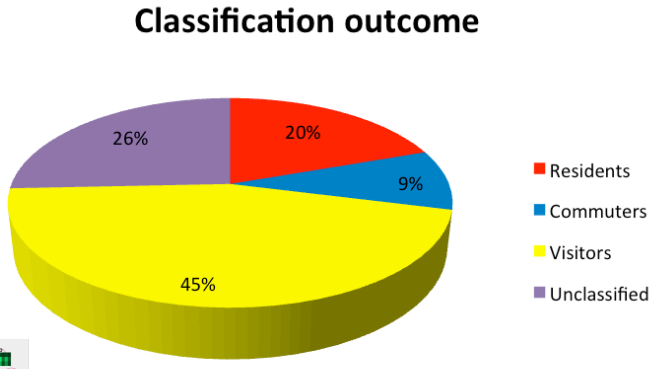
In Commuters

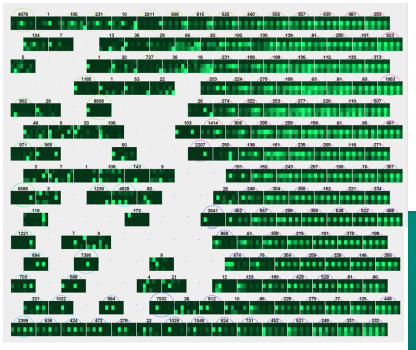


Visitors



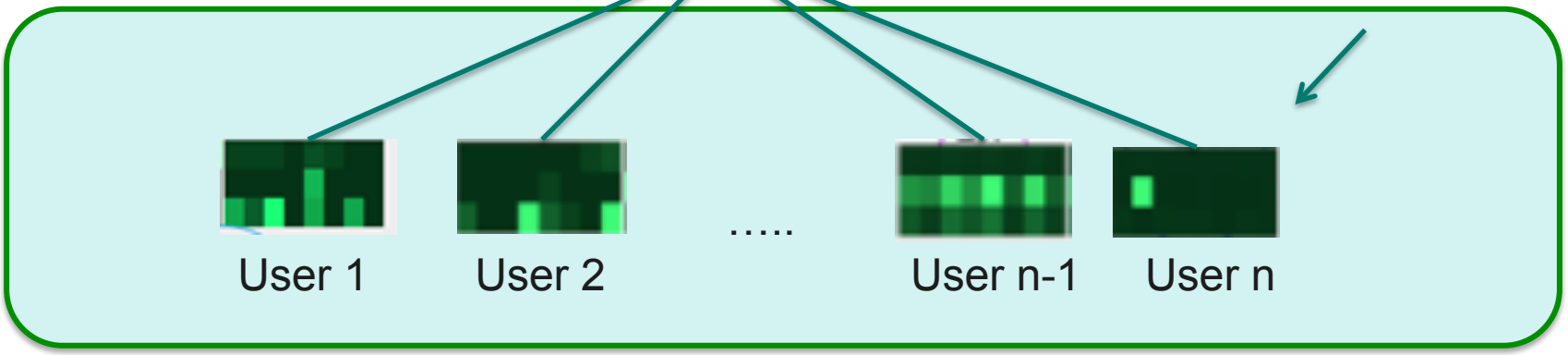
SOCIO-METER FRAMEWORK



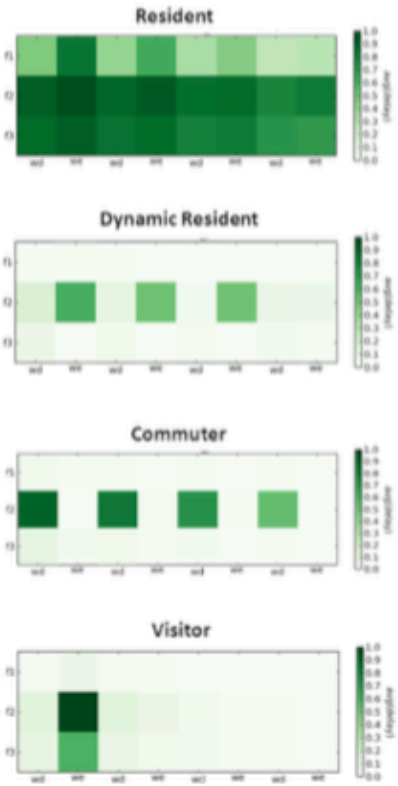


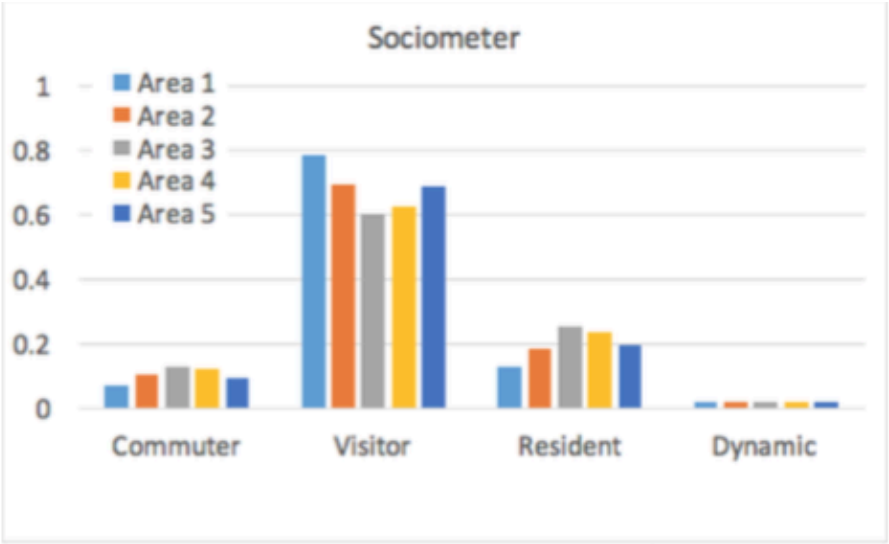
Clustering for Building profiles

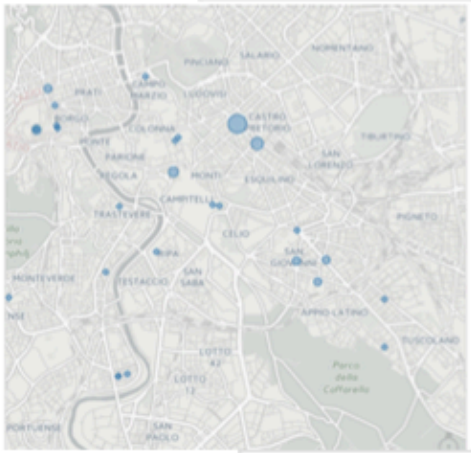
Aggregated call activities computed by the Telco Operator

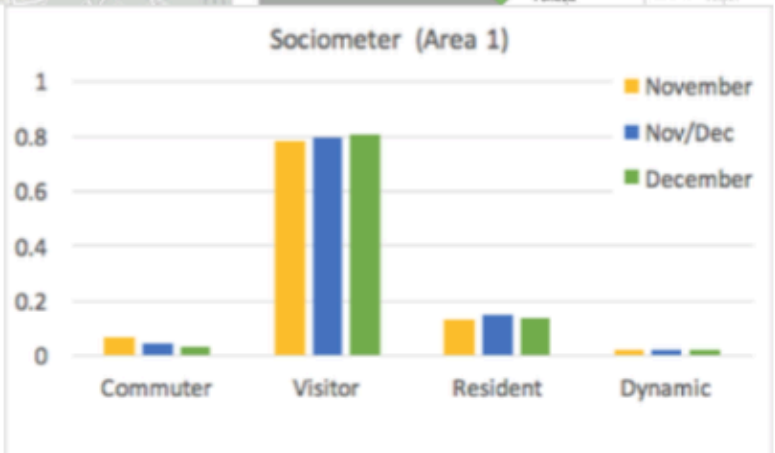


REAL TIME DEMOGRAPHY

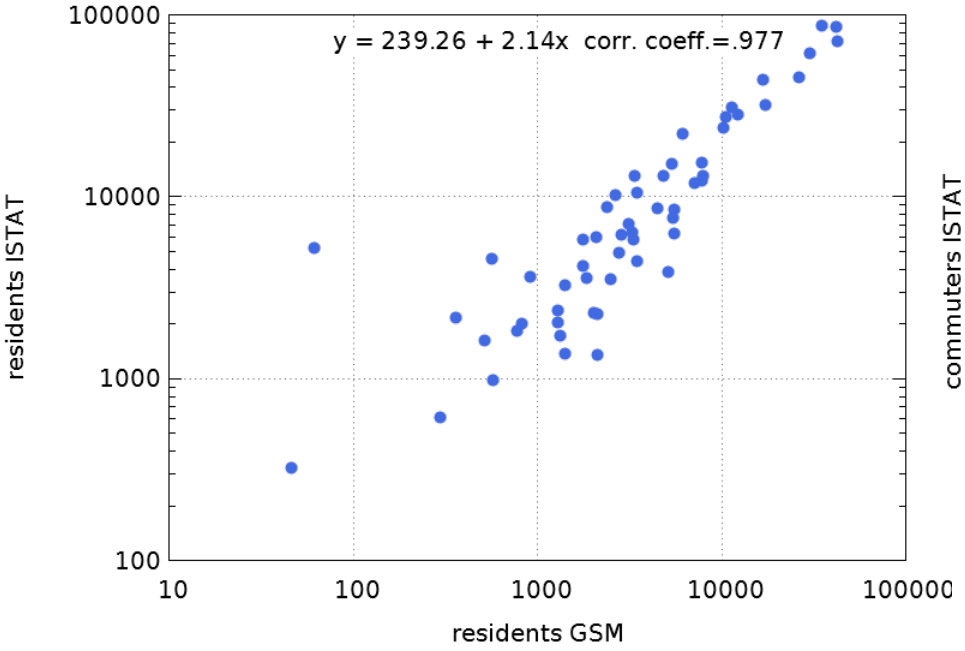


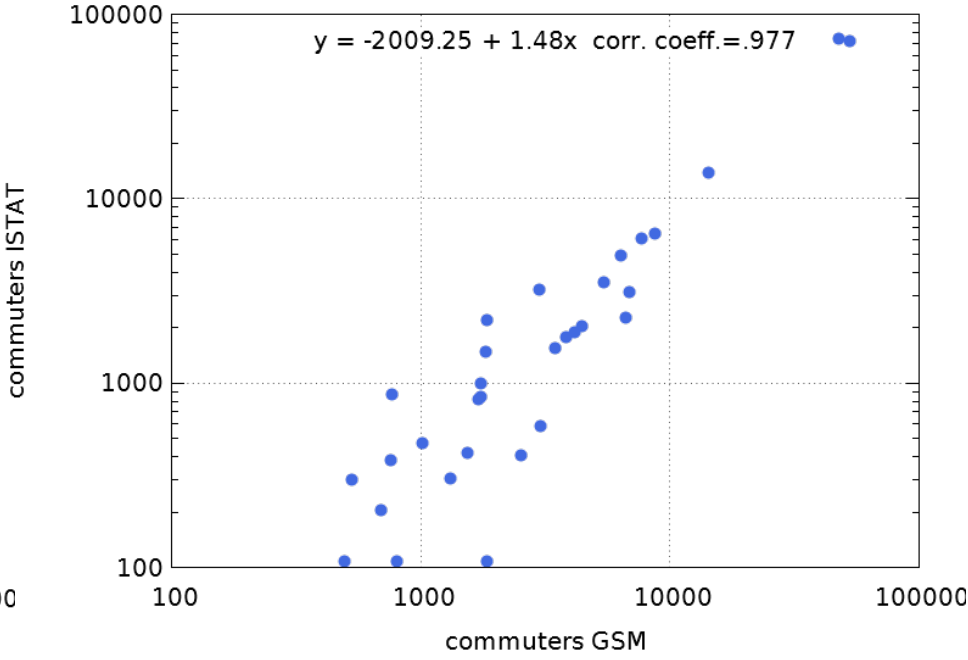


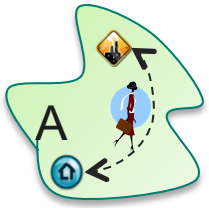


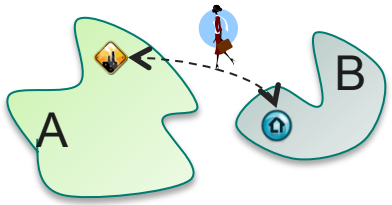


VALIDATION WITH ADMINISTRATIVE DATA









Join work with ISTAT: Barbara Furletti, Lorenzo Gabrielli, Giuseppe Garofalo, Fosca Giannotti, Letizia Milli, Mirco Nanni, Dino Pedreschi, Roberta Vivio. Use of mobile phone data to estimate mobility flows. Measuring urban population and intercity mobility using big data in an integrated approach. Italian Symposium on Statistics (2014).

Big Data Analytics & Social Mining

a tool to

measure,

understand,

and possibly predict

human behavior

Data Scientist needs to take into account ethical and legal aspects and social impact of data science

DATA MINING TASKS...

- Clustering
- Classification
- Pattern Mining & Association Rules

SO, WHAT IS DATA?

Collection of data objects and their attributes

An attribute is a property or characteristic of an object

- Examples: eye color of a person, temperature, etc.
- Attribute is also known as variable, field, characteristic, or feature

A collection of attributes describe an object

- Object is also known as record, point, case, sample, entity, or instance

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

SUPERVISED VS. UNSUPERVISED LEARNING

Supervised learning (classification)

- Supervision: The training data (observations, measurements, etc.) are accompanied by **labels** indicating the class of the observations
- New data is classified based on the training set

Unsupervised learning (clustering)

- The class labels of training data is unknown
- Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

CLUSTERING

CLUSTERING DEFINITION

Cluster: A collection of data objects

Given a set of data points, each having a **set of attributes**, and a **similarity measure** among them, find clusters such that

- Data points in one cluster are more similar to one another.
- Data points in separate clusters are less similar to one another.

Similarity Measures?

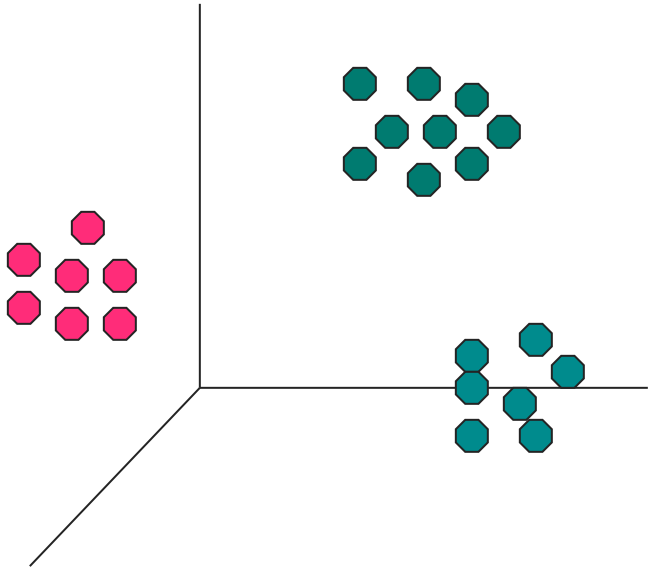
- Euclidean Distance if attributes are continuous.
- Other Problem-specific Measures.

ILLUSTRATING CLUSTERING

Euclidean Distance Based Clustering in 3-D space.

Intracluster distances are minimized

Intercluster distances are maximized

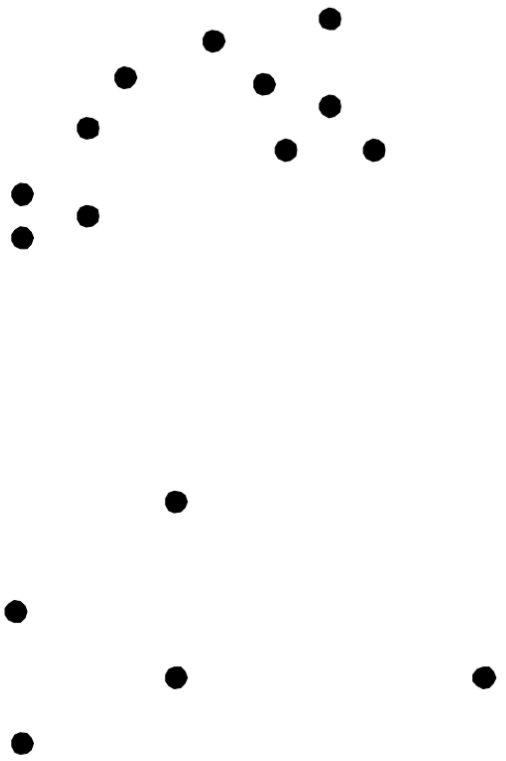


DIFFERENT CLUSTERING APPROACHES

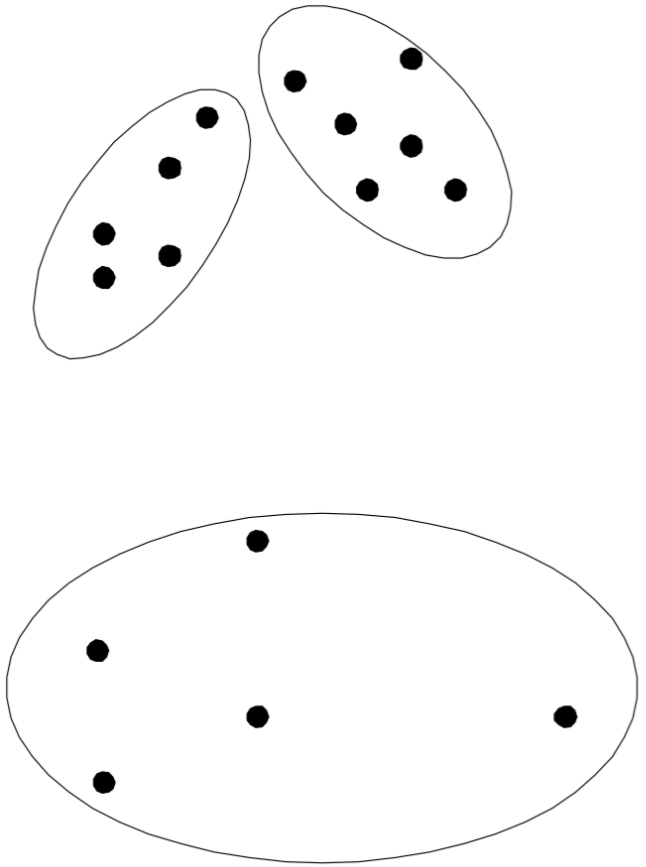
PARTITIONING ALGORITHMS
Directly divides data points into some prespecified number of clusters without a hierarchical structure

HIERARCHICAL ALGORITHMS
Groups data with a sequence of nested partitions, either from singleton clusters to a cluster containing all elements, or viceversa

PARTITIONING CLUSTERING



Original Points



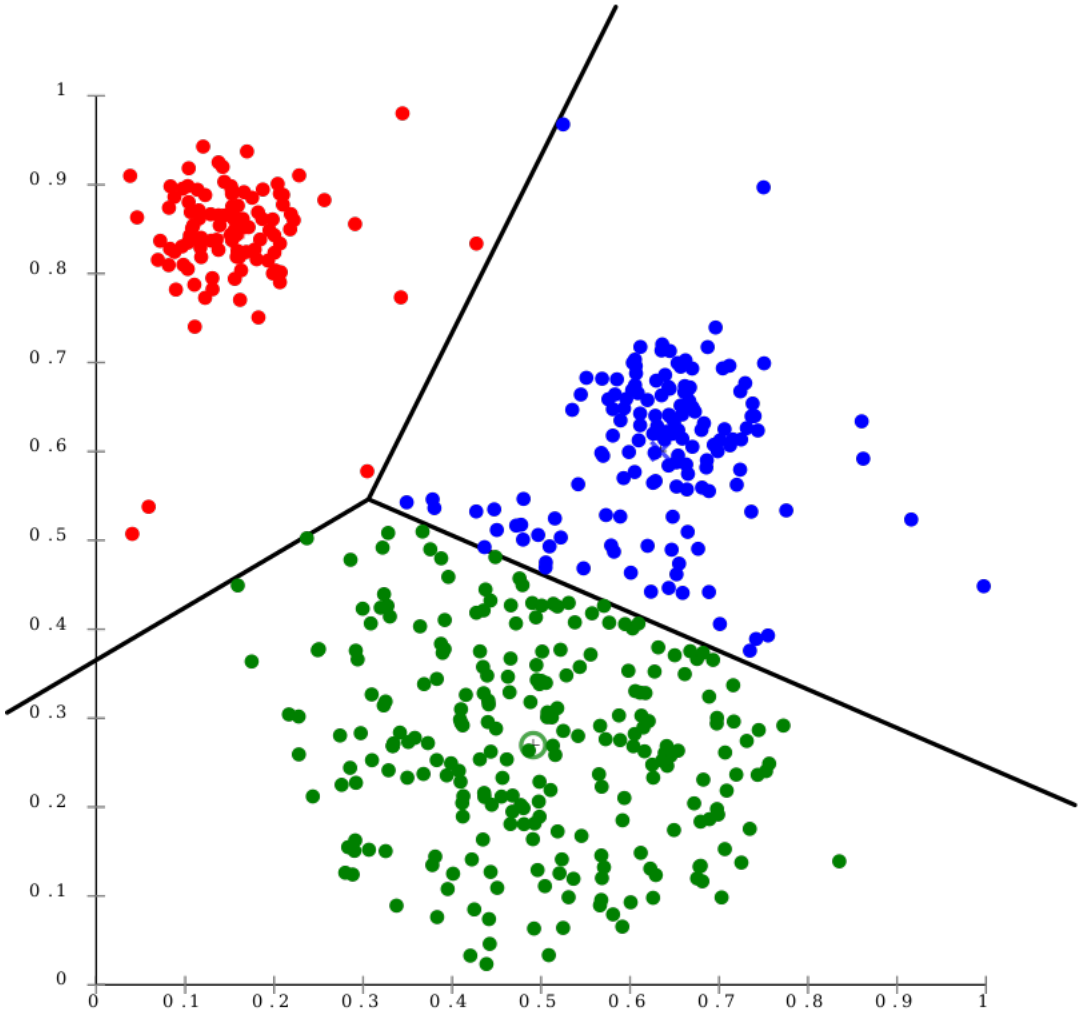
A Partitional Clustering

CENTER-BASED CLUSTERING

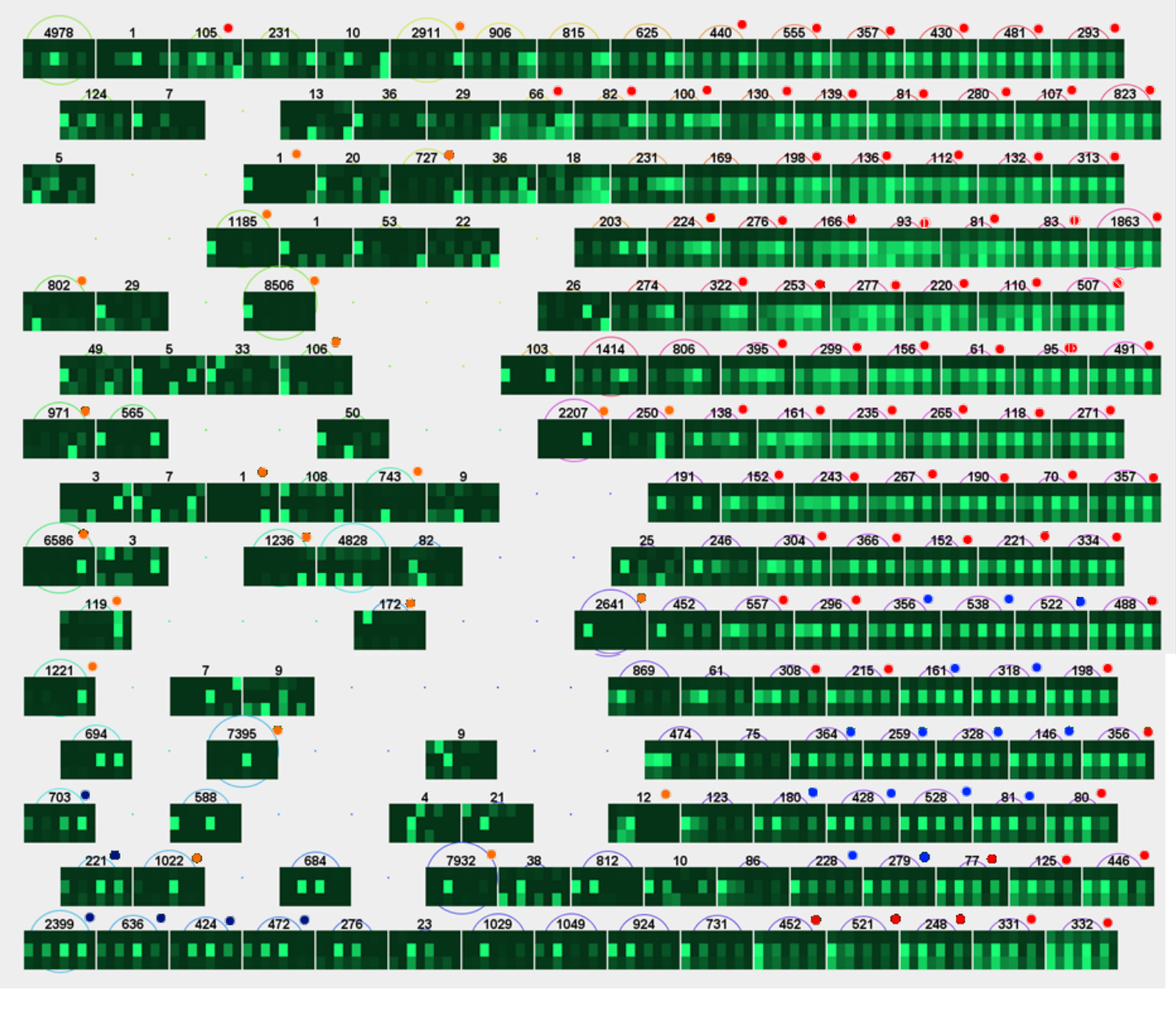
A cluster is a set of objects such that an object in a cluster is **closer (more similar) to the "center"** of a cluster, than to the center of any other cluster

The center of a cluster is often a **centroid**, the average of all the points in the cluster, or a **medoid**, the most "representative" point of a cluster

K-MEANS OR K-MEDOID



CLUSTERING: APPLICATION 1 - CENTROID



d call
nputed
lco
or

CLUSTERING: APPLICATION 2 - CENTROID

Market Segmentation

Goal: subdivide a market into distinct **subsets of customers** where any subset may conceivably be selected as a market target to be reached with a **distinct marketing mix.**

Approach

1. **Collect different attributes** of customers based on their geographical, **Demographic,** lifestyle, **Behavioral** related information
2. **Find clusters** of similar customers
3. Measure **the clustering quality** by observing buying patterns of customers in same cluster vs. those from different clusters.

A BEHAVIOR BASED SEGMENTATION EXAMPLE

Using unsupervised clustering segmentation for a grocery chain which would like better product assortment for its high profitable customers

Potential Inputs

Value	• Basket Size • Visit Frequency
Basket	• Spend by category • Type of category • Brand spend (i.e. private label)
Promotions	• % bought on targeted promotion • % bought from flyer
Time	• Time of day • Day of week
Location	• Store format • Area population density

Clustering approach

Deal Seeking Mom

Key Differentiators

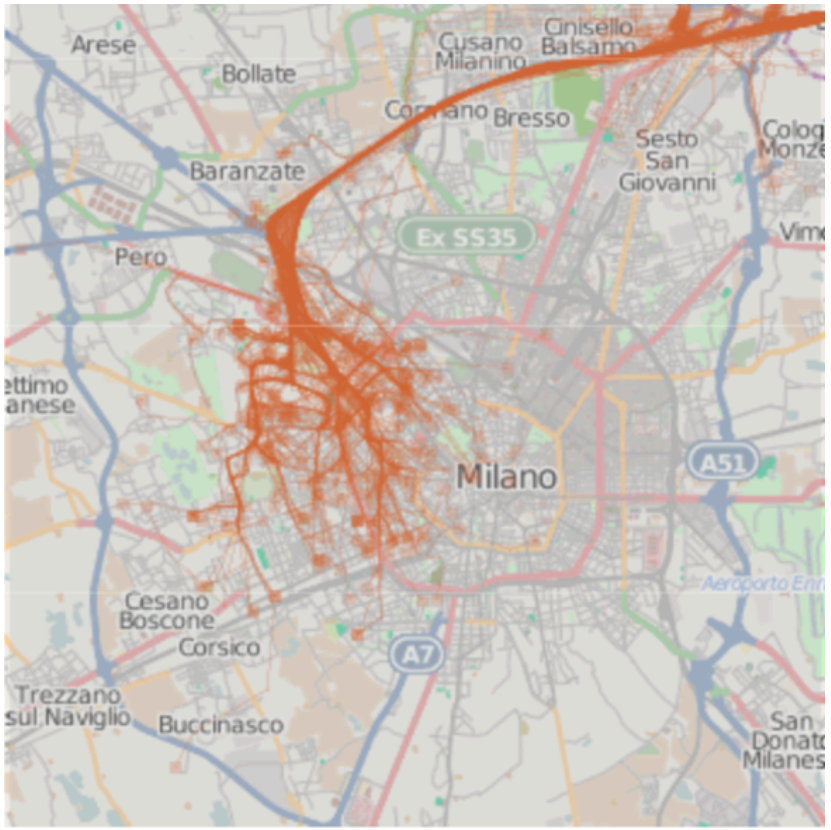
- Full store shop
- High avg. basket size / # trips

- High % purchased on promotion
- Rewards seeker

- High spend categories
  - Fresh produce
  - Organic food
  - Multipack juice, snack

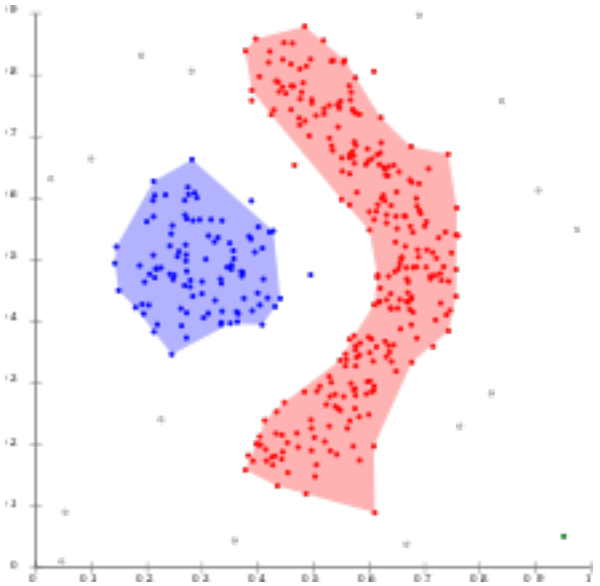
A PARTICULAR APPLICATION: TRAJECTORIES

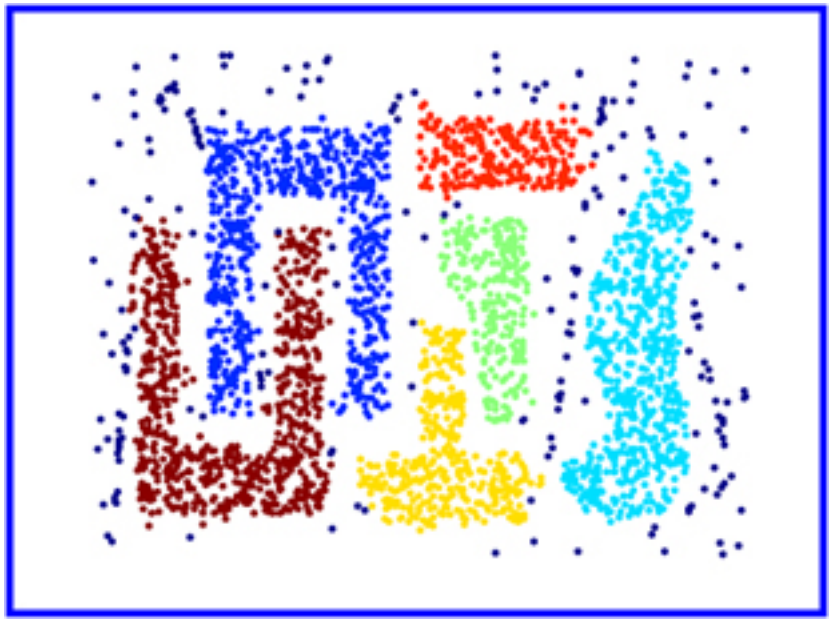
NO GLOBULAR CLUSTERS



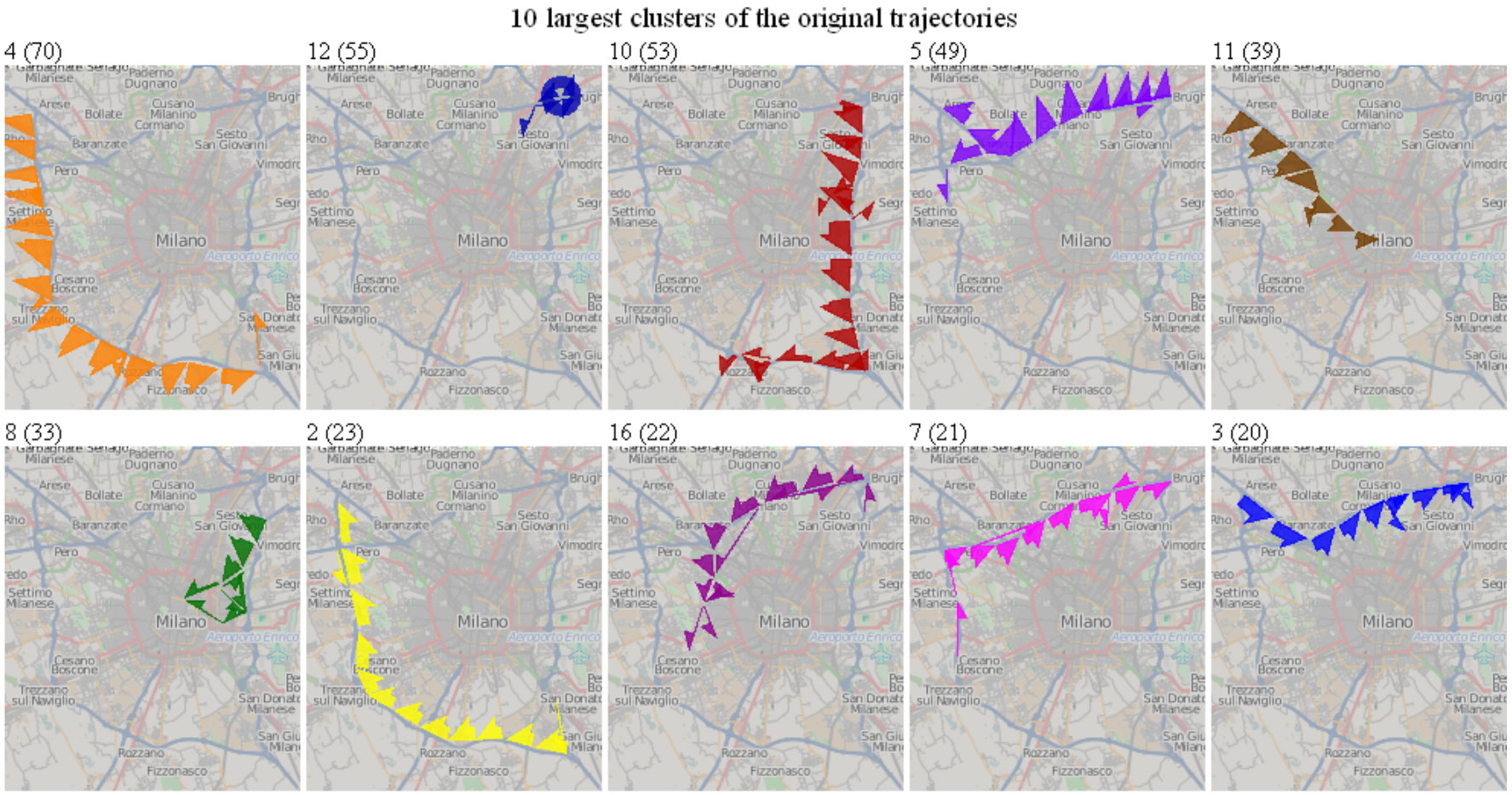
DENSITY-BASED CLUSTERING

Clusters are **dense regions** in the data space seprated by regions with **lower density**

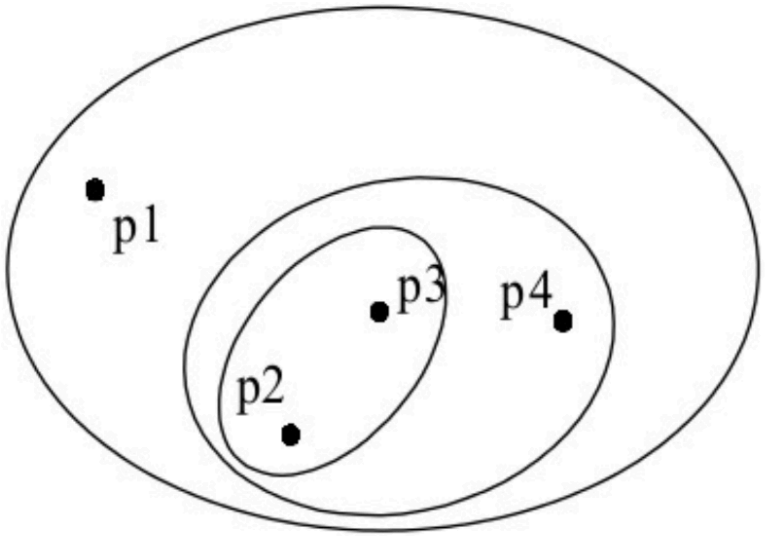




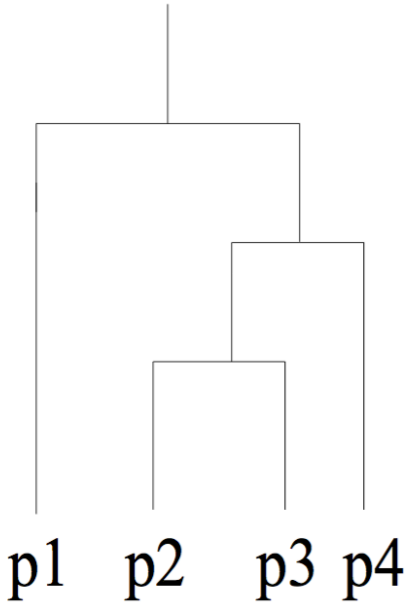
CLUSTERING OF TRAJECTORIES



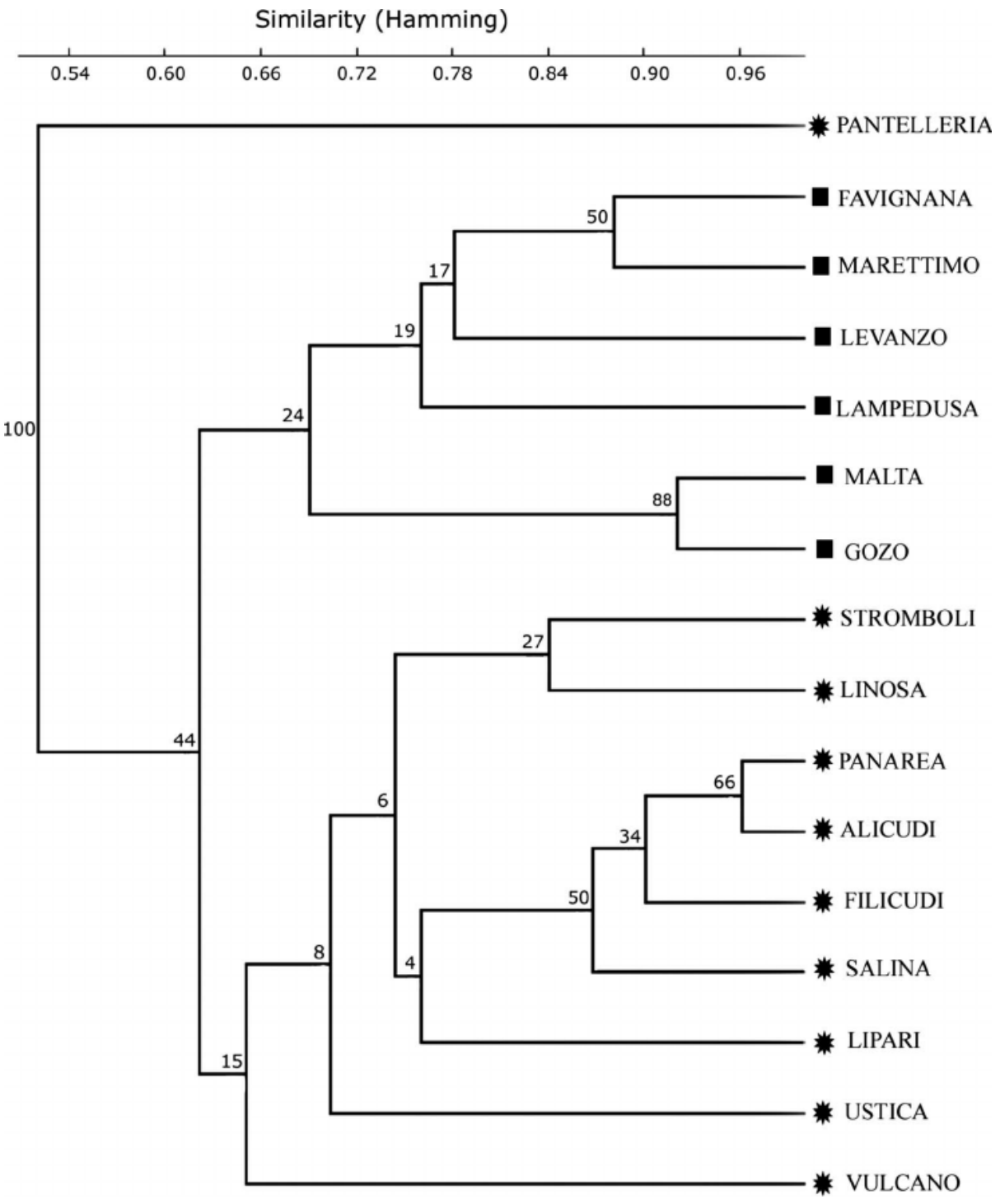
HIERARCHICAL CLUSTERING



Hierarchical Clustering



Dendrogram


Similarity (Hamming)
0.54
0.60
0.66
0.72
0.78
0.84
0.90
0.96
PANTELLERIA
FAVIGNANA
MARETTIMO
LEVANZO
LAMPEDUSA
MALTA
GOZO
STROMBOLI
LINOSA
PANAREA
ALICUDI
FILICUDI
SALINA
LIPARI
USTICA
VULCANO
100
44
24
19
17
50
88
27
6
66
34
50
4
8
15

CLASSIFICATION - PREDICTION

CLASSIFICATION: DEFINITION

Given a collection of records (*training set*)

- Each record contains a set of *attributes*, one of the attributes is the *class*.

Find a *model* for class attribute as a function of the values of other attributes.

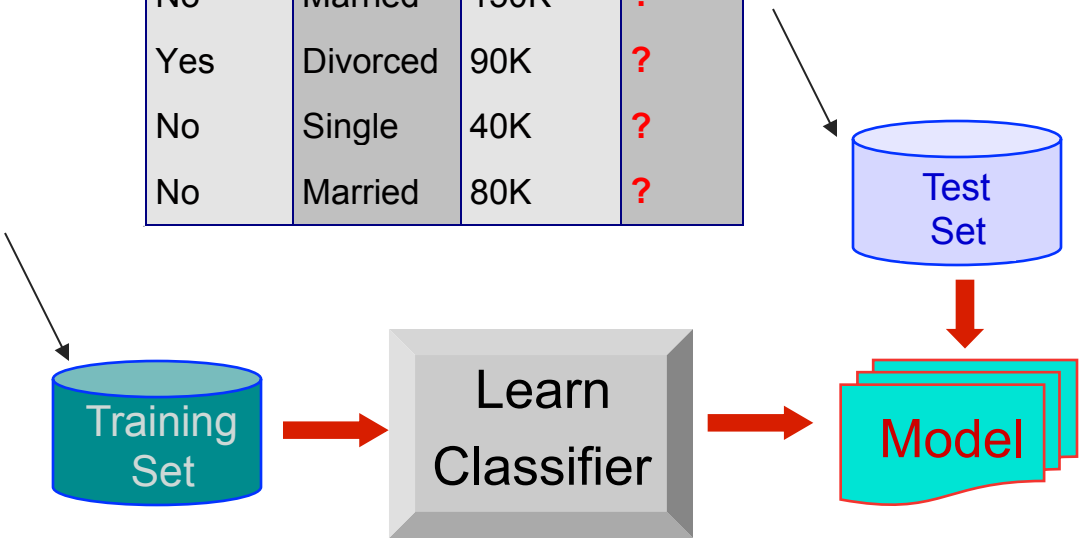
Goal: previously unseen records should be assigned a class as accurately as possible.

- A *test set* is used to determine the accuracy of the model. Usually, the given data set is divided into training and test sets, with training set used to build the model and test set used to validate it.

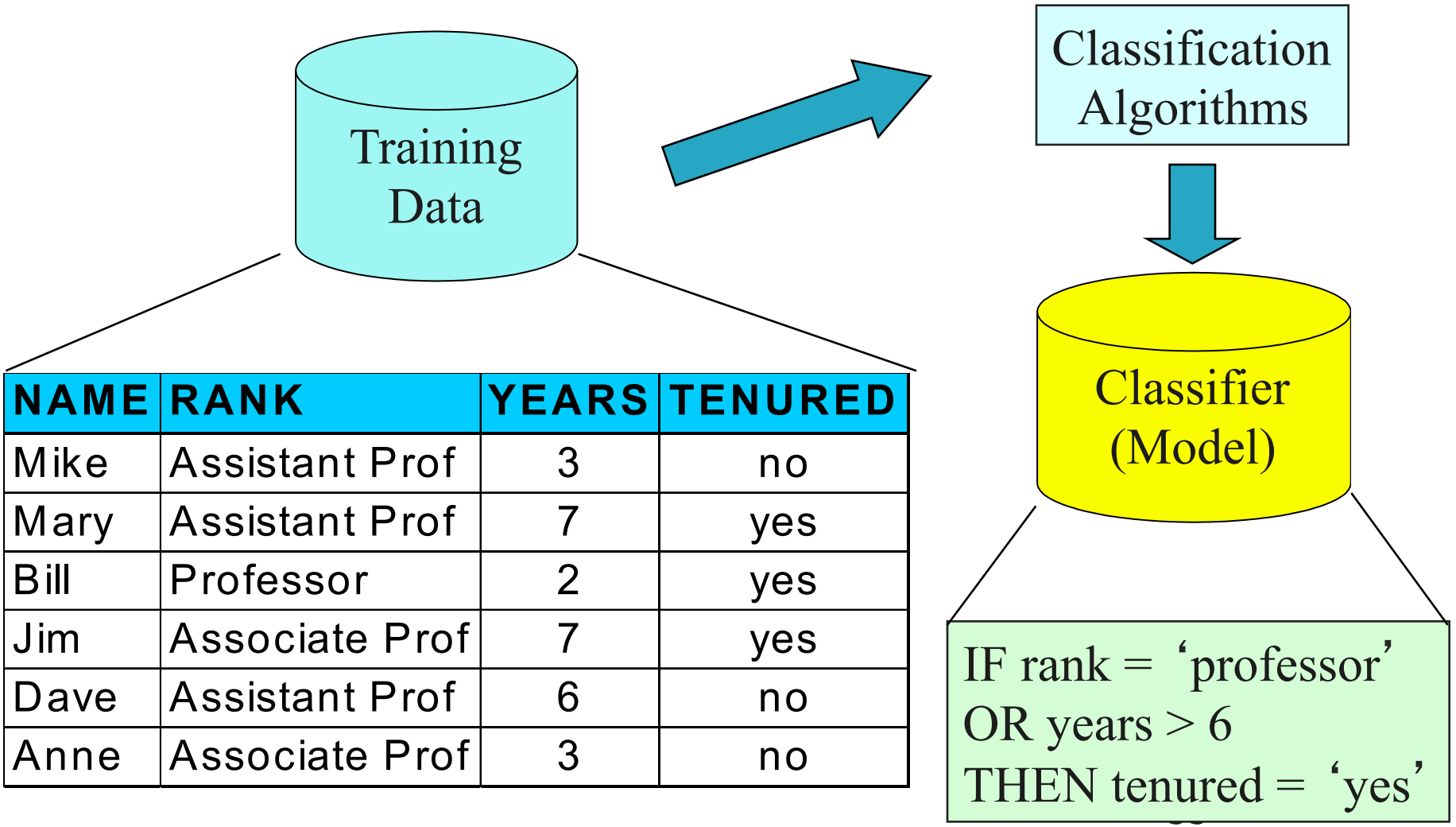
CLASSIFICATION EXAMPLE

Tid	Refund	Marital Status	Taxable Income	Cheat
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75K	?
Yes	Married	50K	?
No	Married	150K	?
Yes	Divorced	90K	?
No	Single	40K	?
No	Married	80K	?

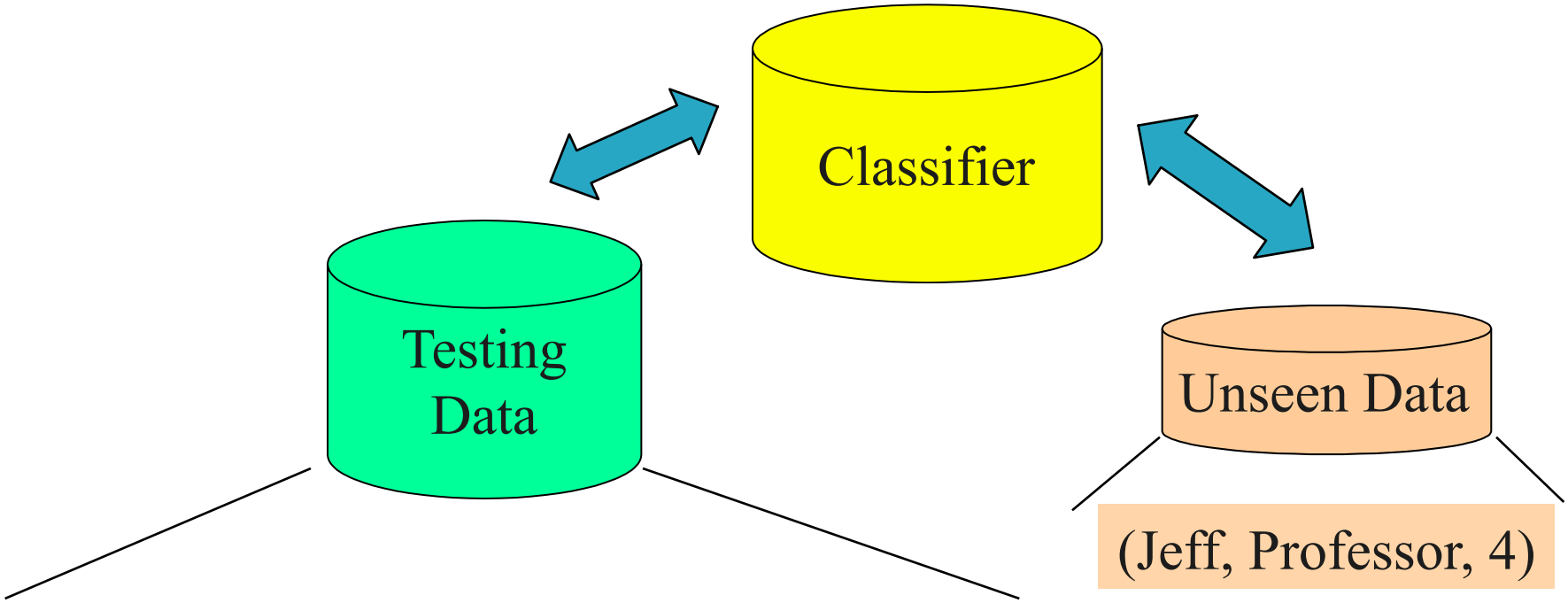


PROCESS (1): MODEL CONSTRUCTION



NAME	RANK	YEARS	TENURED
Mike	Assistant Prof	3	no
Mary	Assistant Prof	7	yes
Bill	Professor	2	yes
Jim	Associate Prof	7	yes
Dave	Assistant Prof	6	no
Anne	Associate Prof	3	no

PROCESS (2): USING THE MODEL IN PREDICTION

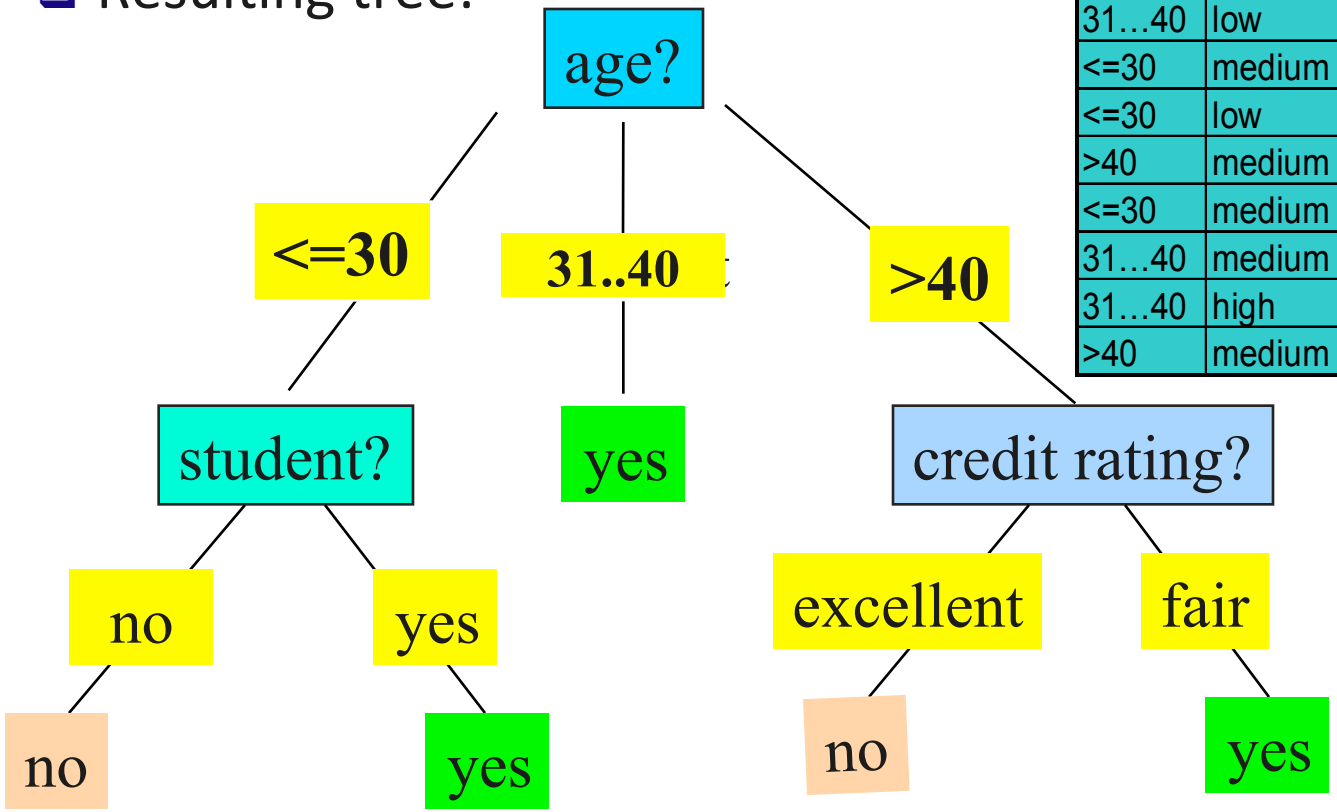


NAME	RANK	YEARS	TENURED
Tom	Assistant Prof	2	no
Merlisa	Associate Prof	7	no
George	Professor	5	yes
Joseph	Assistant Prof	7	yes

DECISION TREE INDUCTION: AN EXAMPLE

- Training data set: Buys_computer
- The data set follows an example of Quinlan's ID3 (Playing Tennis)
- Resulting tree:

age	income	student	credit_rating	buys_computer
<=30	high	no	fair	no
<=30	high	no	excellent	no
31…40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31…40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31…40	medium	no	excellent	yes
31…40	high	yes	fair	yes
>40	medium	no	excellent	no



CLASSIFICATION: APPLICATION

Fraud Detection

Goal: Predict fraudulent cases in credit card transactions.

Approach: Use credit card transactions and the information on its account-holder as attributes.

- When does a customer buy, what does he buy, how often he pays on time, etc

- Label past transactions as fraud or fair transactions. This forms the class attribute.
- Learn a model for the class of the transactions.
- Use this model to detect fraud by observing credit card transactions on an account.

CLASSIFICATION: APPLICATION 2

Customer Attrition/Churn

Goal: To predict whether a customer is likely to be lost to a competitor.

Approach:

1. Use detailed record of transactions with each of the past and present customers, to **find attributes**
   - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, …

2. Label the customers as loyal or disloyal

3. Find a model for loyalty

FREQUENT PATTERN MINING

FREQUENT PATTERN MINING

Determine what items often go together (usually in transactional databases)

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Often Referred to as Market Basket Analysis

- used in retail for planning arrangement on shelves
- used for identifying cross-selling opportunities
- "should" be used to determine best link structure for a Web site

FREQUENT PATTERN MINING

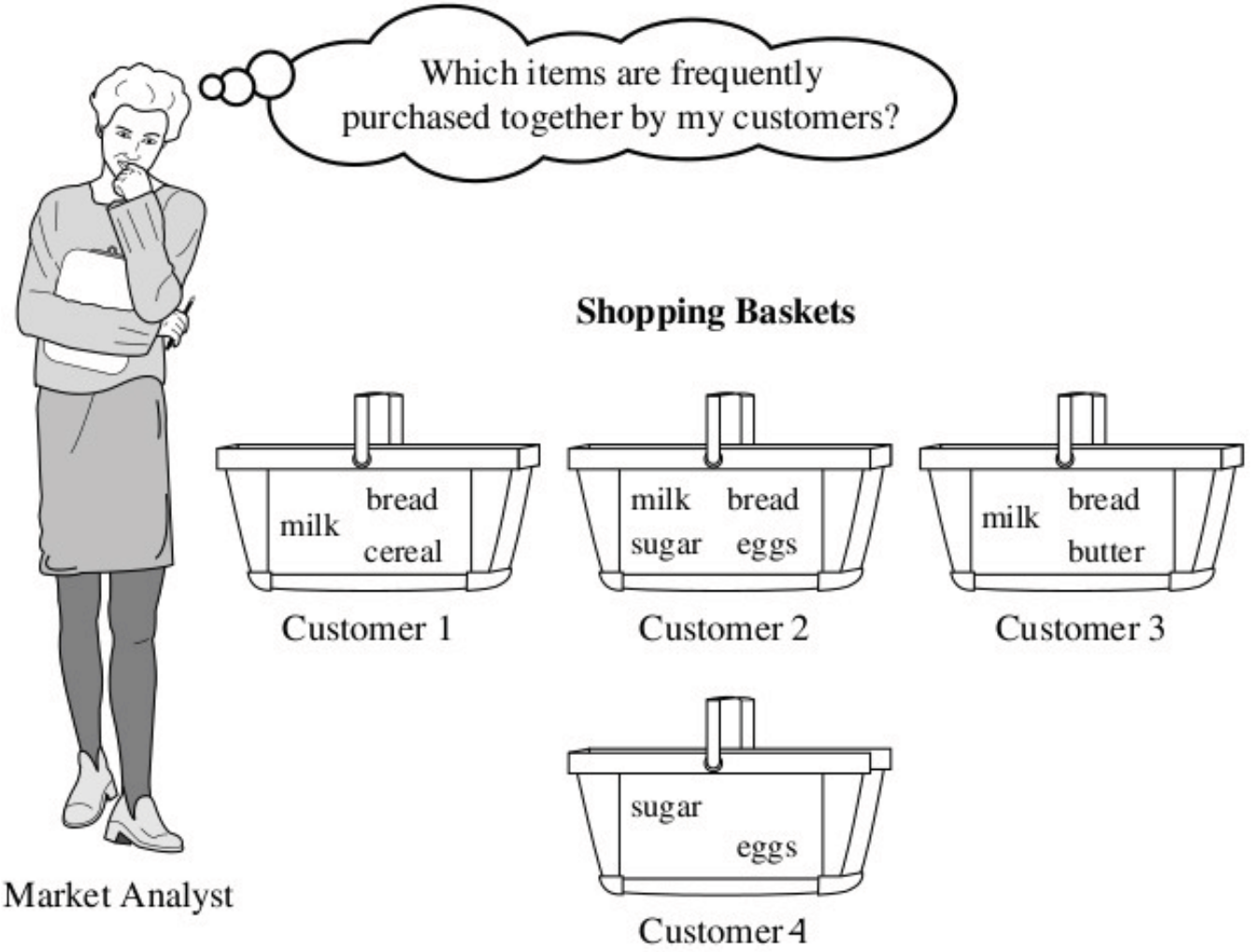
A **Frequent pattern** is a pattern (a set of items, subsequences, subgraphs, etc.) that occurs frequently in a data set.

Motivation

Finding inherent regularities (associations) in data.

Examples

- people who buy milk and beer also tend to buy diapers
- people who access pages A and B are likely to place an online order


Which items are frequently purchased together by my customers?
Shopping Baskets
milk
bread
cereal
Customer 1
milk
bread
sugar
eggs
Customer 2
milk
bread
butter
Customer 3
sugar
eggs
Customer 4
Market Analyst

ASSOCIATION RULE DISCOVERY

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:

{Milk} --> {Coke}

{Diaper, Milk} --> {Beer}

ASSOCIATION RULE DISCOVERY: APPLICATION 1

Marketing and Sales Promotion:

Let the rule discovered be

{Bagels, … } --> {Potato Chips}

Potato Chips as consequent => Can be used to determine what should be done to boost its sales.

Bagels in the antecedent => can be used to see which products would be affected if the store discontinues selling bagels.

Bagels in antecedent *and* Potato chips in consequent => Can be used to see what products should be sold with Bagels to promote sale of Potato chips!

MATERIAL

Learning Material

- **Textbook:** Pang-Ning Tan, Michael Steinbach, Vipin Kumar **Introduction to DATA MINING** Addison Wesley, ISBN 0-321-32136-7, 2006
- **Textbook**: Berthold, M.R., Borgelt, C., Höppner, F., Klawonn, F**. GUIDE TO INTELLIGENT DATA ANALYSIS.** Springer Verlag, 1st Edition., 2010. ISBN 978-1-84882-259-7

Slides of the classes

Data mining software

- **KNIME:** The Konstanz Information Miner. Download page