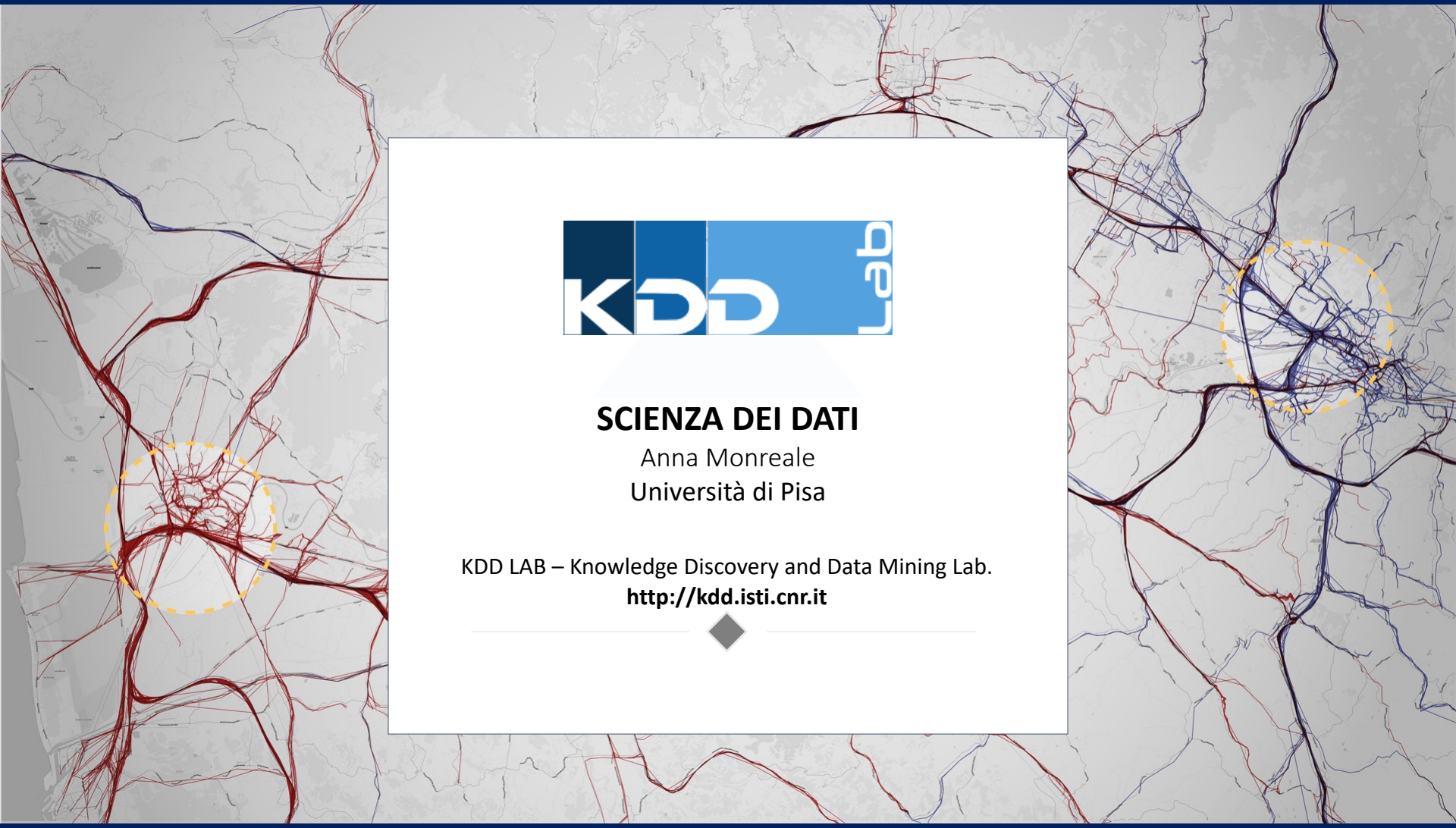


SCIENZA DEI DATI

Anna Monreale
Università di Pisa

KDD LAB – Knowledge Discovery and Data Mining Lab.
http://kdd.isti.cnr.it

INTRODUCTION

Big Data Analytics & Social Mining

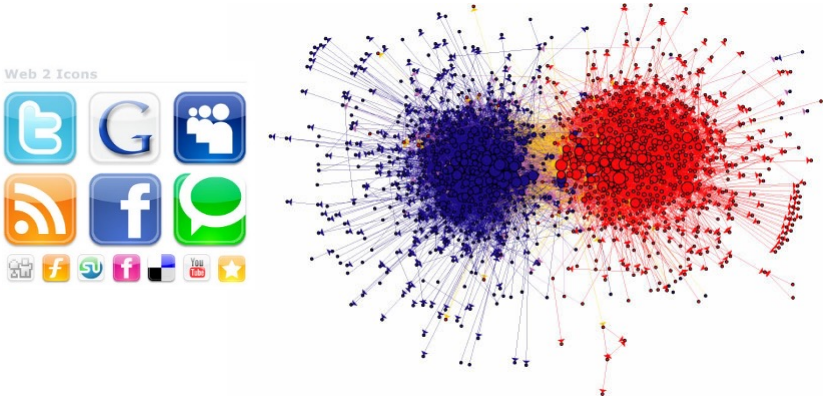

2005
Luca Bruno / AP


2013
NBC NEWS
Michael Sohn / AP

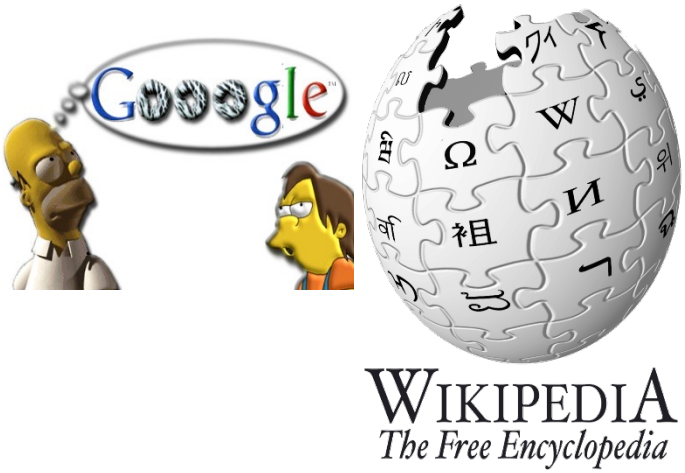
Big data proxies della vita sociale

Shopping patterns & lyfestyle

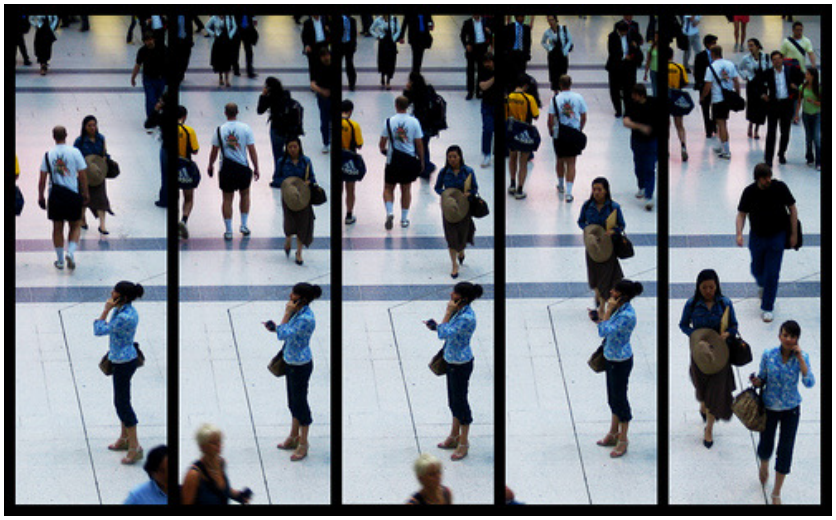
RELATIONSHIPS & SOCIAL TIES



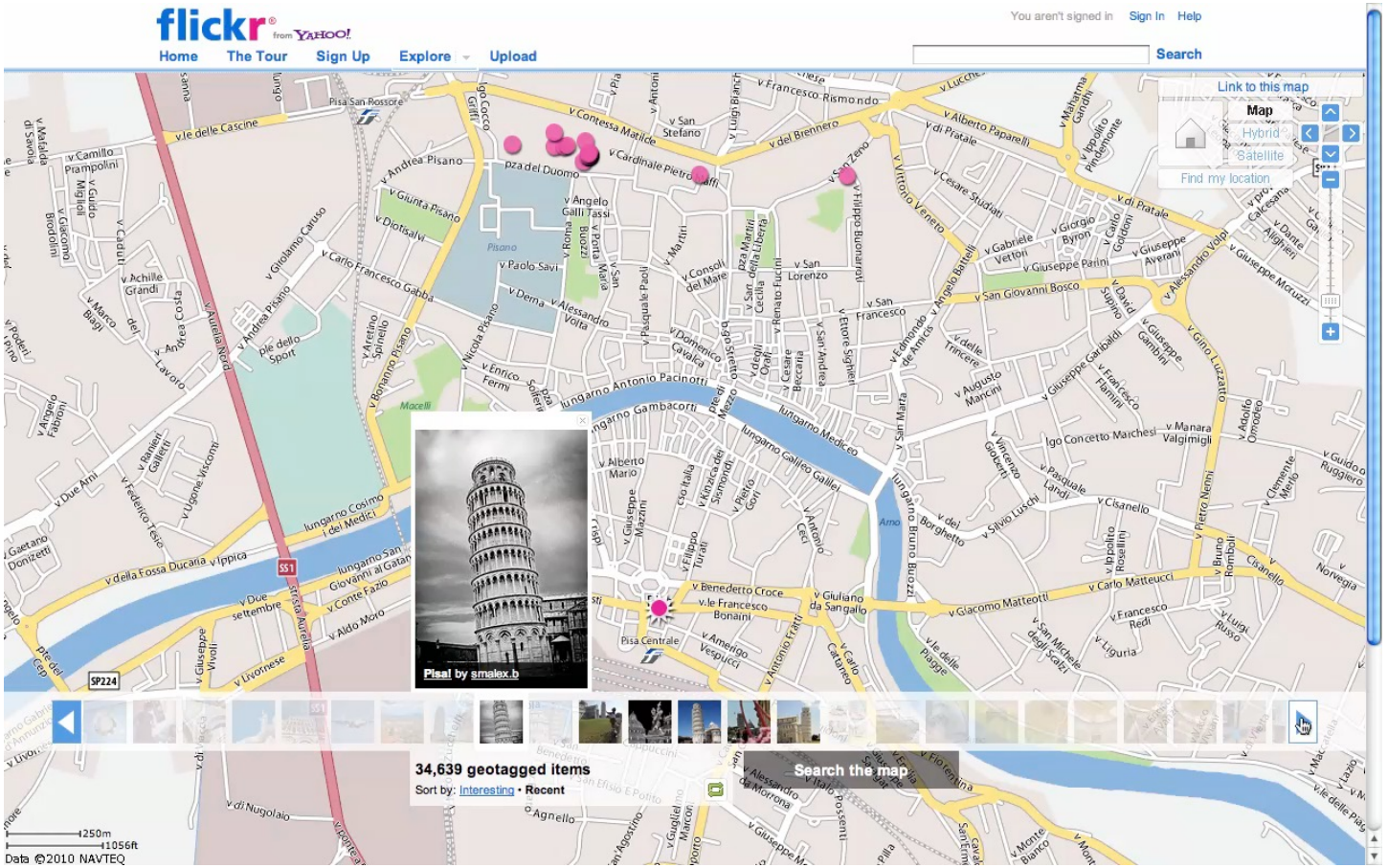
DESIRES, OPINIONS, SENTIMENTS



MOVEMENTS

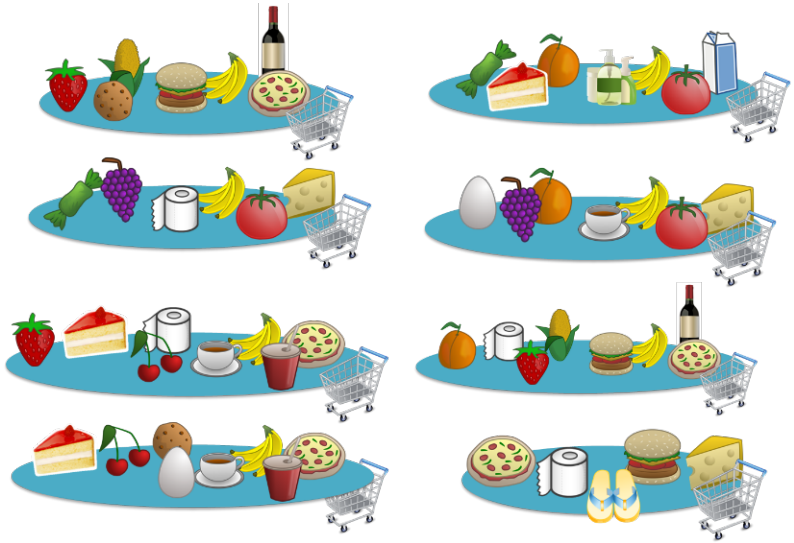


Social Networks



Transazioni economiche

- Dati di spesa dei clienti fidelizzati
  - Storia degli acquisti individuali
  - Non solo quanto il cliente ha acquistato, ma anche quando e cosa…



GPS tracce veicolari

GPS coordinate collezionati/spediti da navigatori e “scatole nere”

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

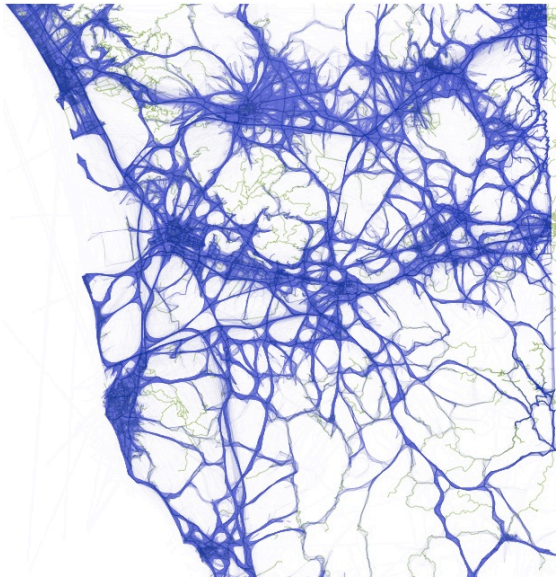
…

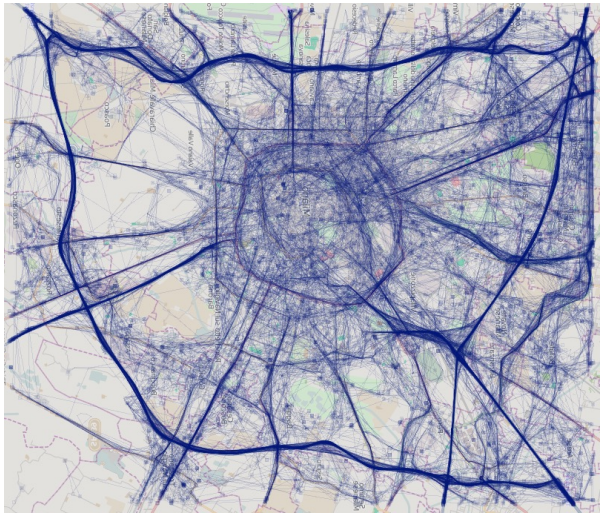
8;22/03/07 08:51:52;50.777132;7.205580; 67.6;345.4;21.817;3.8;1808;4
8;22/03/07 08:51:56;50.777352;7.205435; 68.4;35.6;14.223;3.8;1808;4
8;22/03/07 08:51:59;50.777415;7.205543; 68.3;112.7;25.298;3.8;1808;4
8;22/03/07 08:52:03;50.777317;7.205877; 68.8;119.8;32.447;3.8;1808;4
8;22/03/07 08:52:06;50.777185;7.206202;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

…

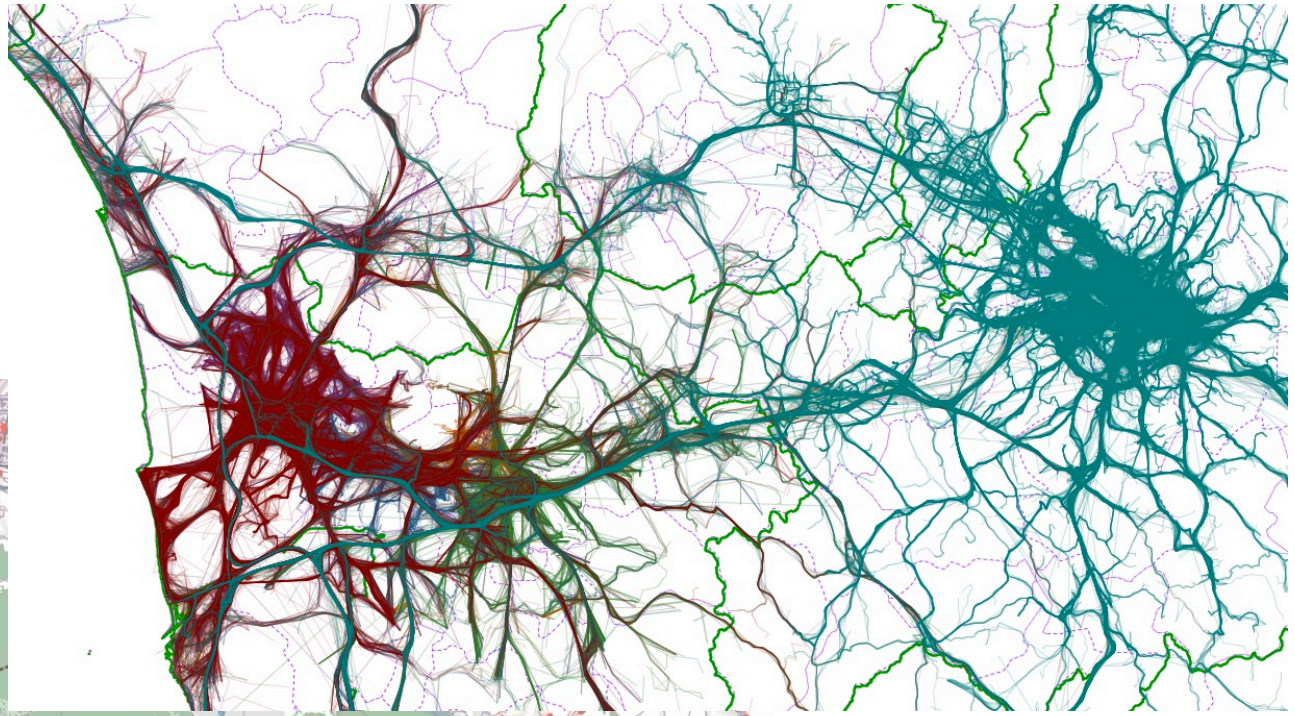
Tempo di raccolta: ~30-60 secs

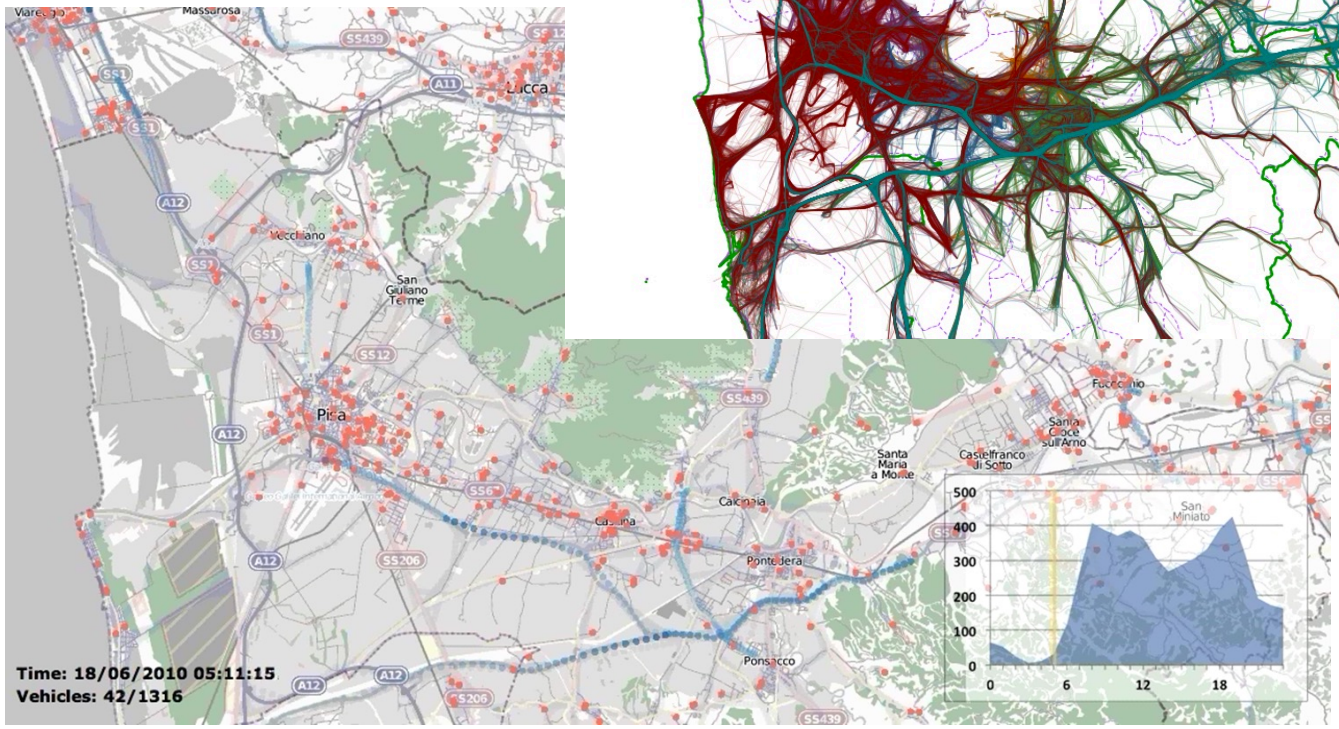
Errore di localizzazione: ~5-10 m



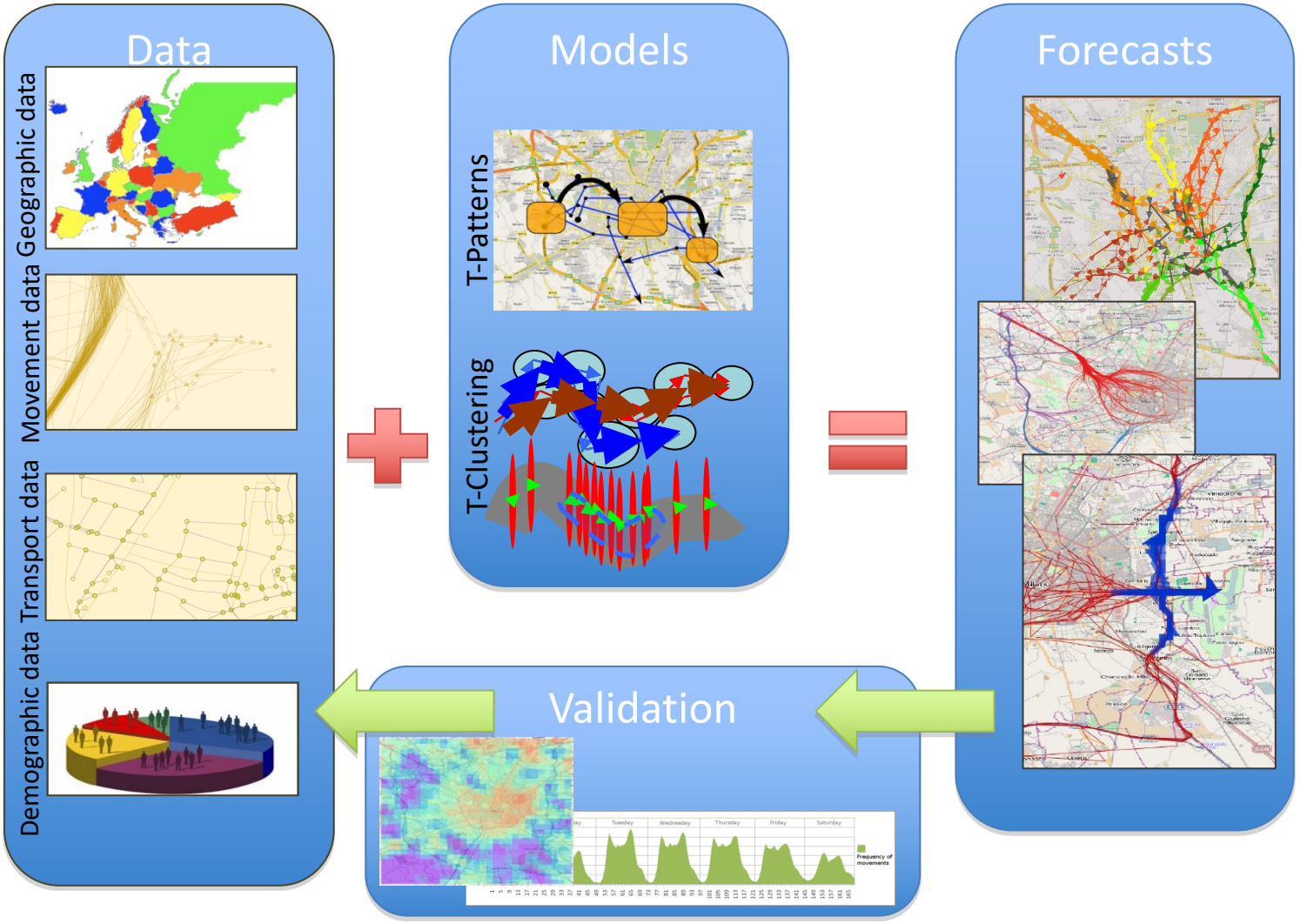


GPS tracce veicolari

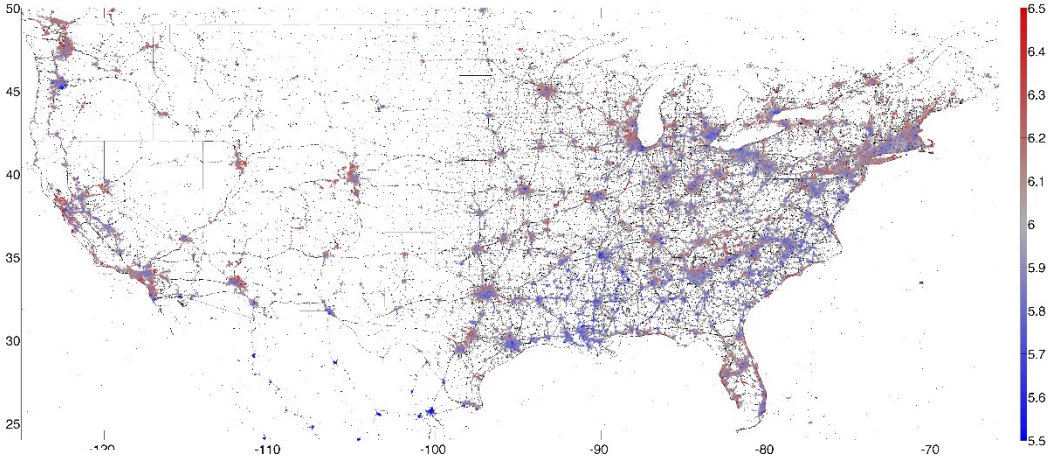


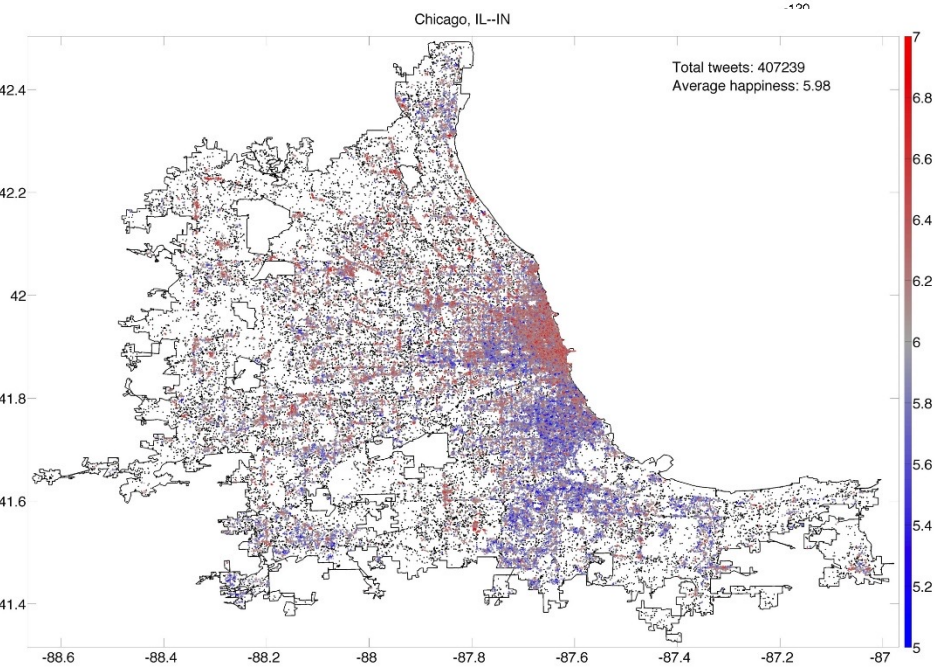


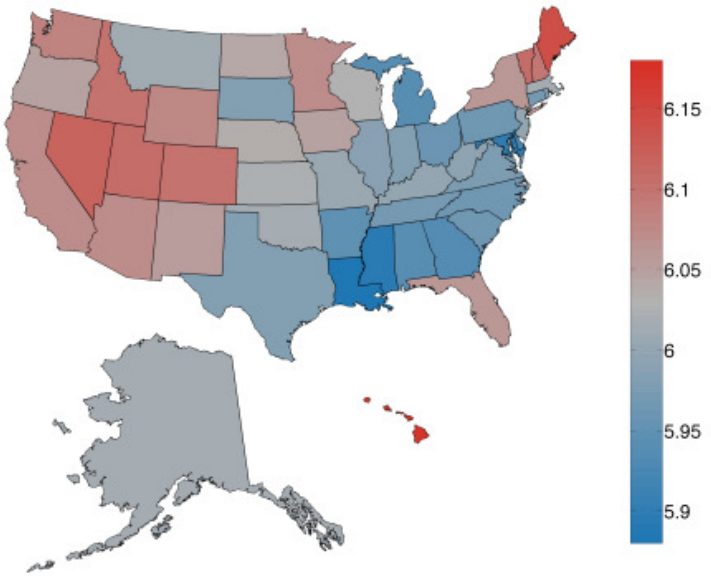
Dai dati alla conoscenza



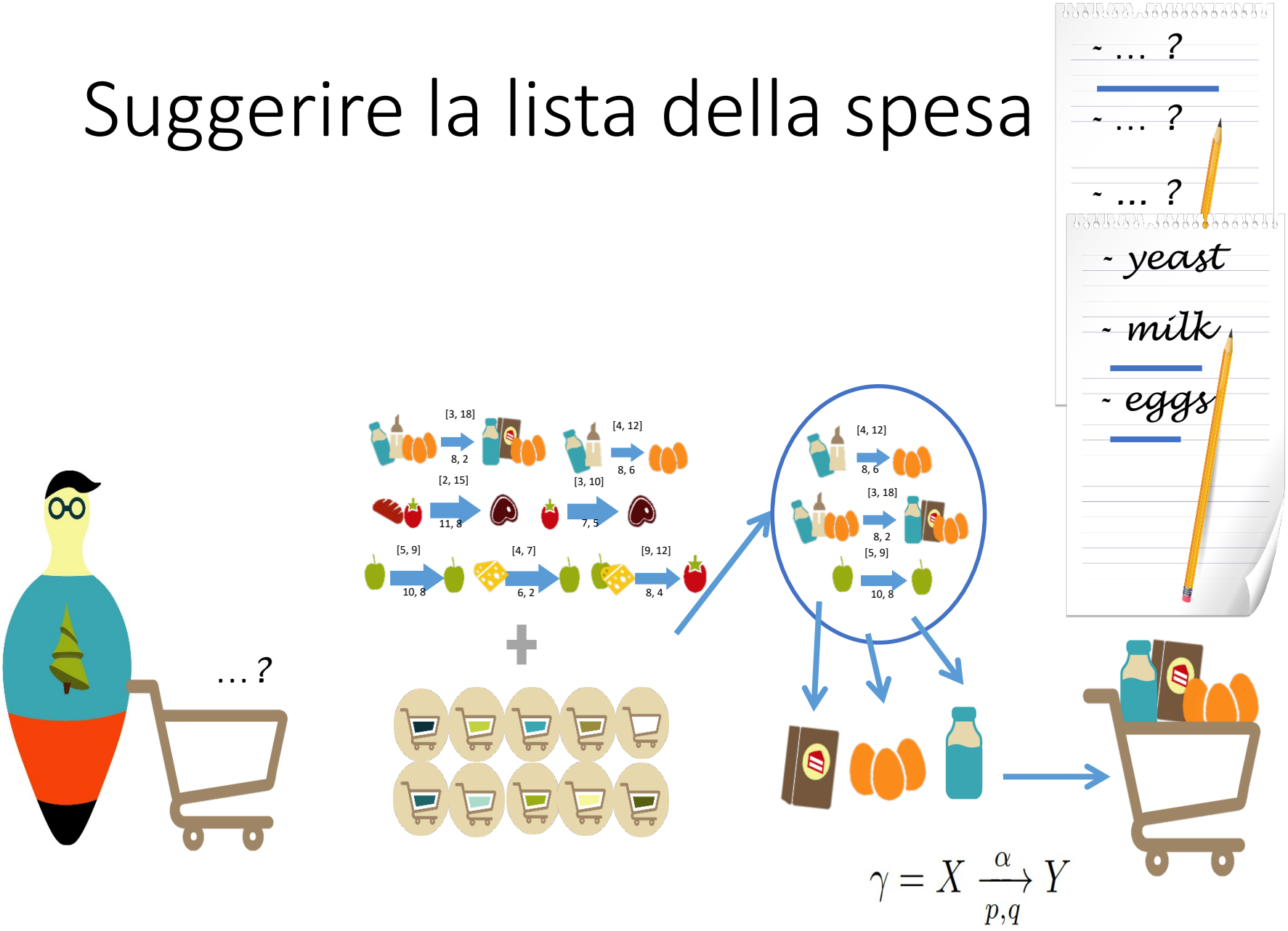
Misurare la felicità attraverso Twitter





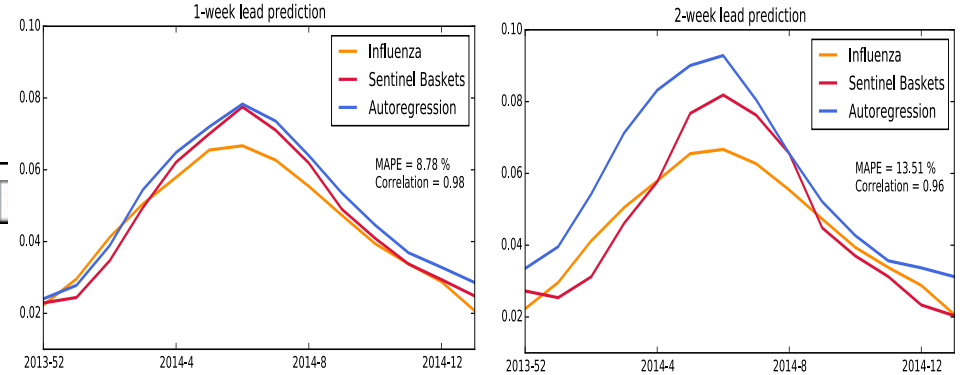


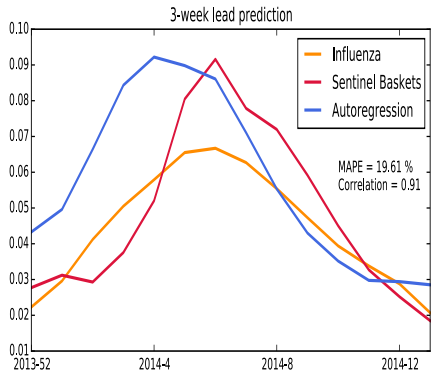
Suggerire la lista della spesa

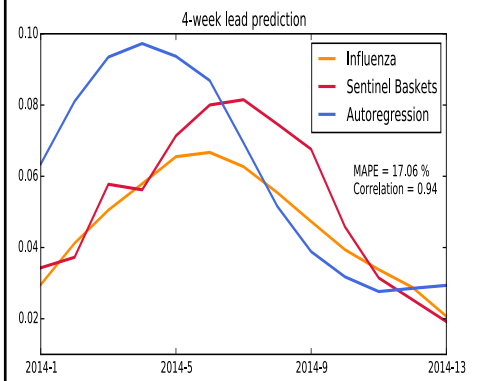


Predire il picco influenzale

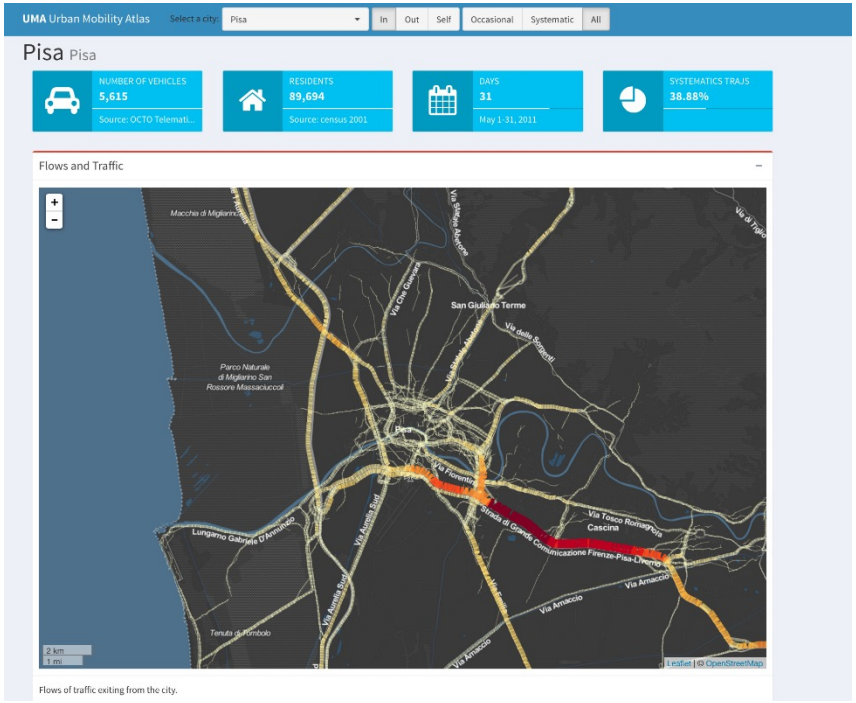
2013/14 Influenza season predicted values (top5):

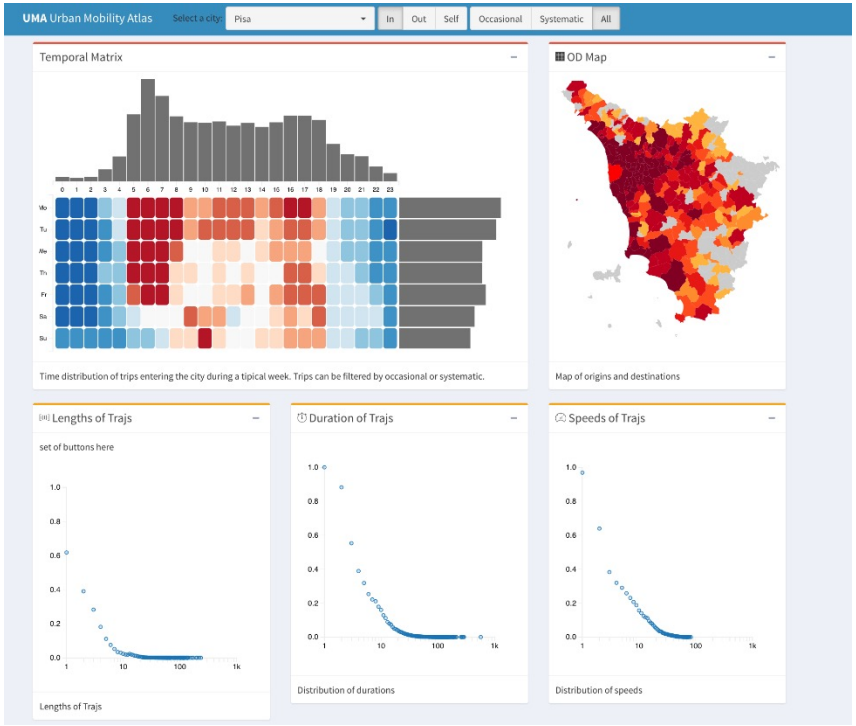






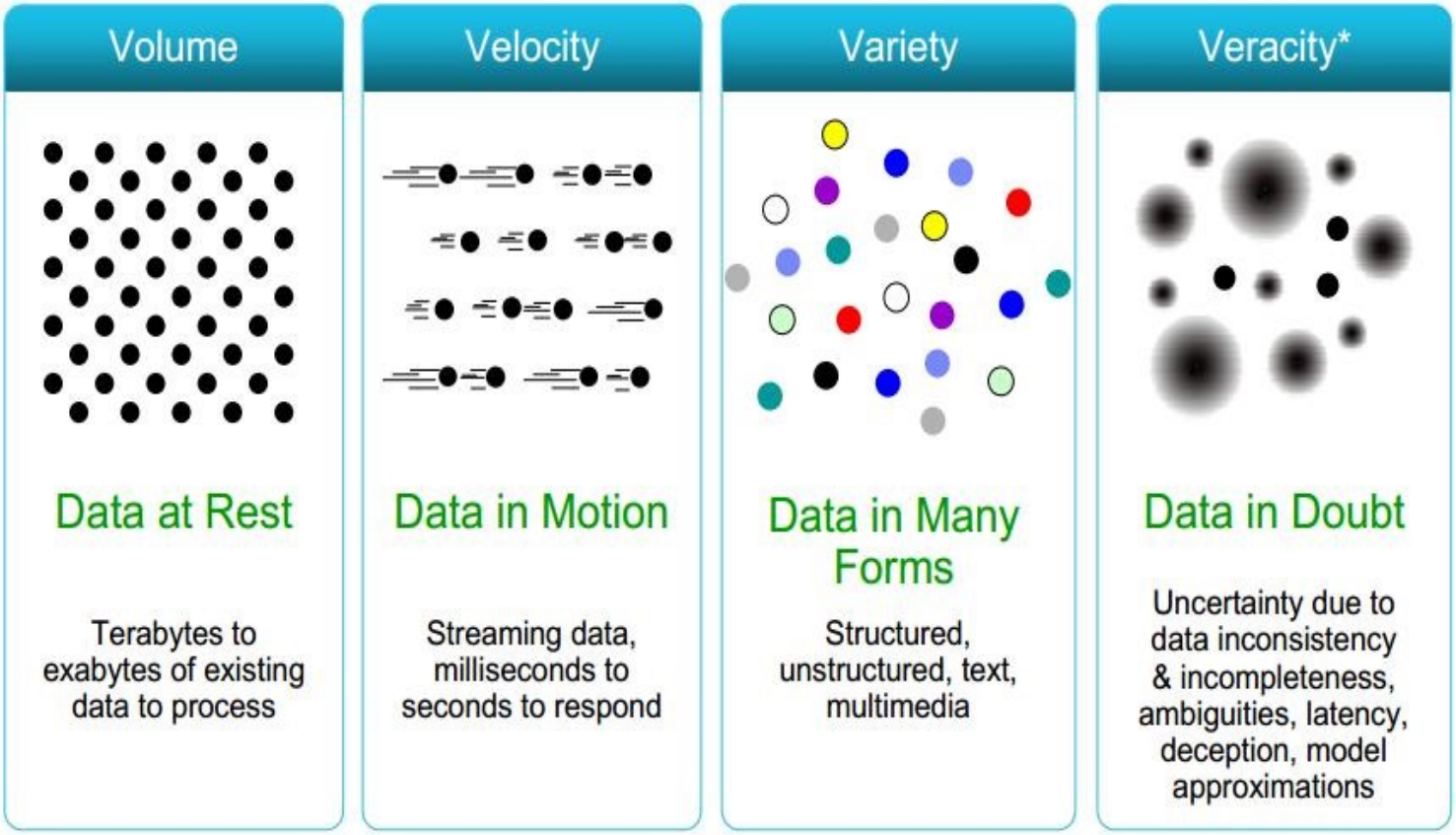
L’atlante della mobilità urbana





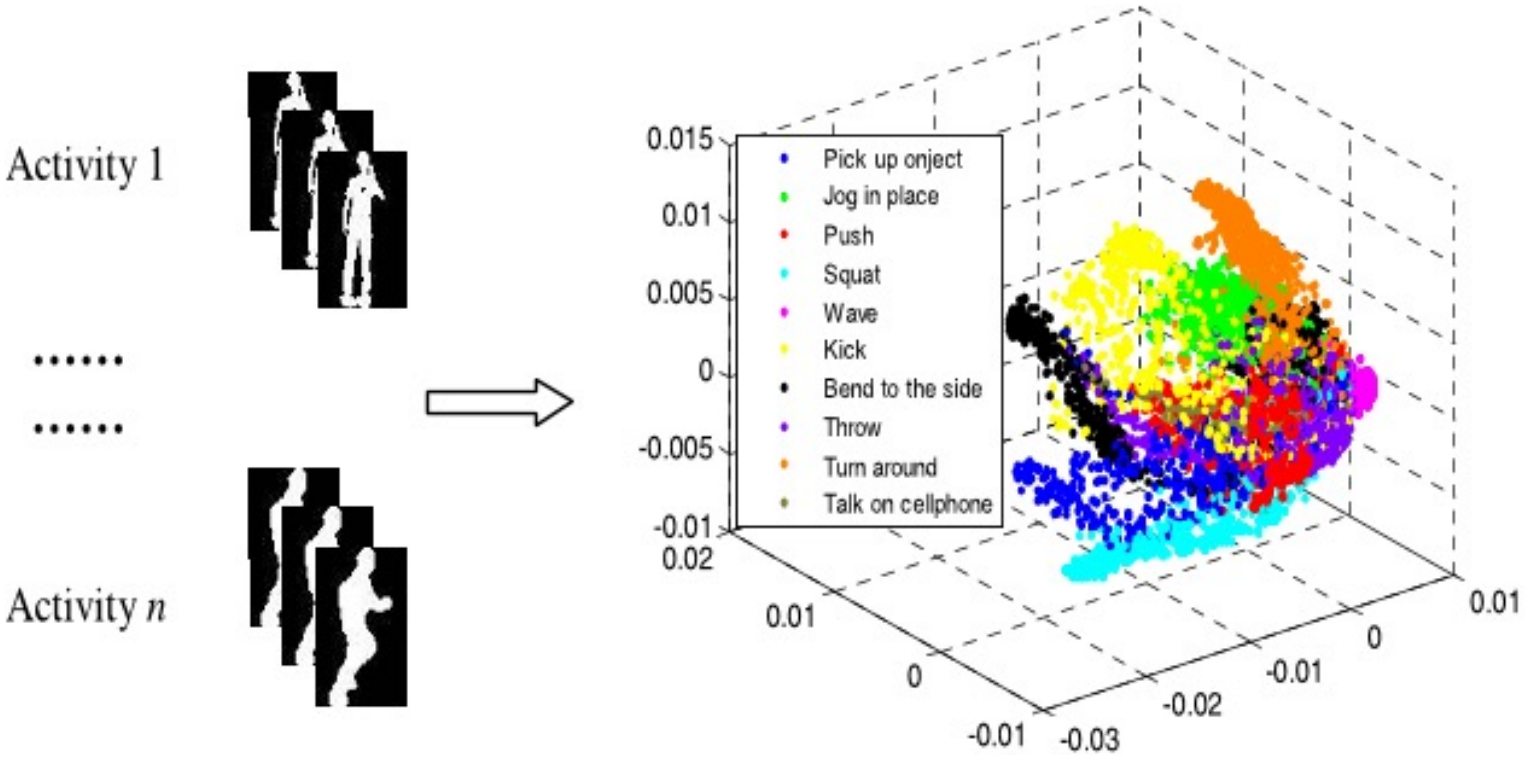
I Big Data e le 4V

Dati possono essere strutturati, non-strutturati, piccoli, grandi, statici, dinamici,…

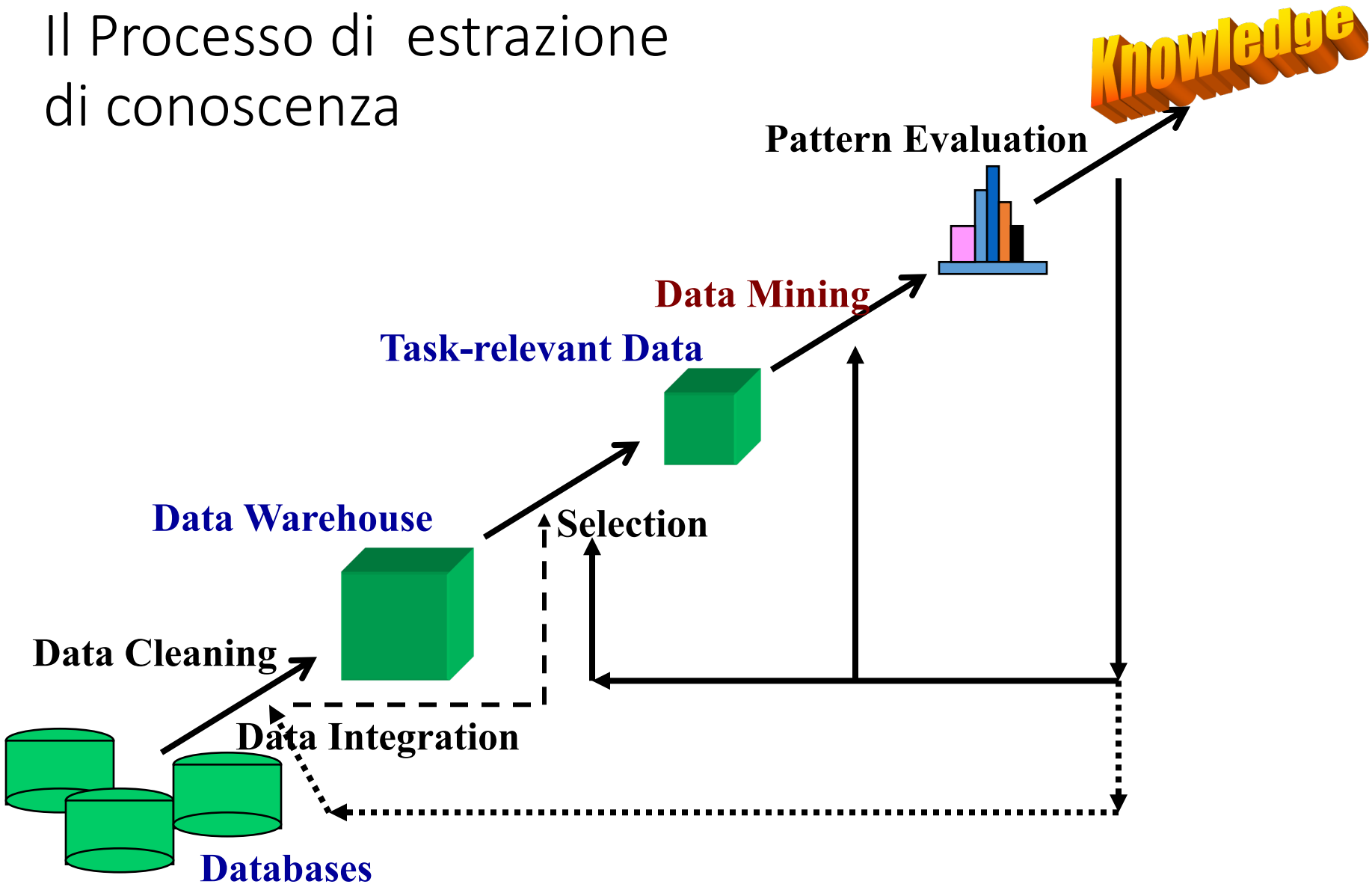


Come scopriamo la conoscenza?

Algoritmi di Data mining per la scoperta automatica di patterns che rivelano la struttura nascosta dentro enormi datasets.



Il Processo di estrazione di conoscenza



Tipi di dati

Dati tabulari

- Ogni oggetto è descritto da proprietà (attribute)
- Un **attributo** è una proprietà o caratteristica dell'oggetto
  - Esempi: colore degli occhi, temperatura, ecc.
- Object è anche conosciuto come, record, punto, caso, esempio, entità, o istanza

Attributi

Oggetti

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

Data Matrix

Projection of x Load	Projection of y load	Distance	Load	Thickness
10.23	5.27	15.22	2.7	1.2
12.65	6.25	16.22	2.2	1.1

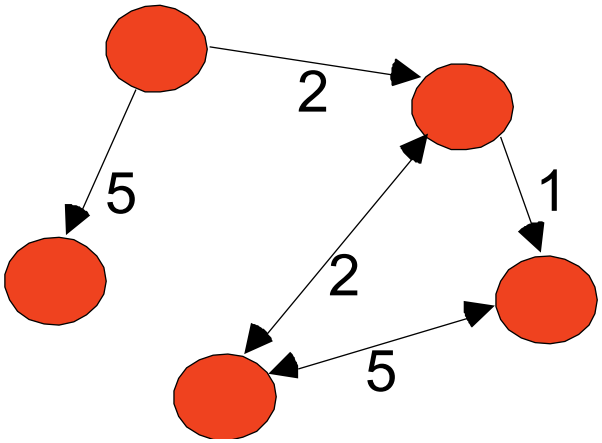
Document Data

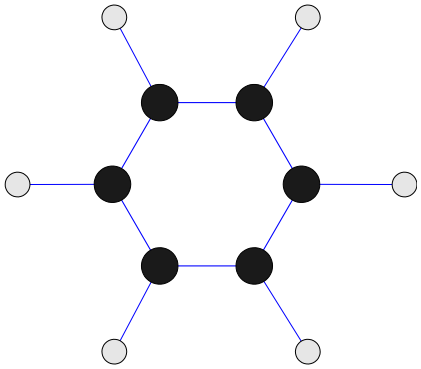
	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Transaction Data

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

Graph Data





Benzene Molecule: C6H6

Useful Links:

- Bibliography
- Other Useful Web sites
  - ACM SIGKDD
  - KDnuggets
  - The Data Mine

Knowledge Discovery and Data Mining Bibliography

(Gets updated frequently, so visit often!)

- Books
- General Data Mining

Book References in Data Mining and Knowledge Discovery

Usama Fayyad, Gregory Piatetsky-Shapiro, Padhraic Smyth, and Ramasamy uthurasamy, "Advances in Knowledge Discovery and Data Mining", AAAI Press/the MIT Press, 1996.

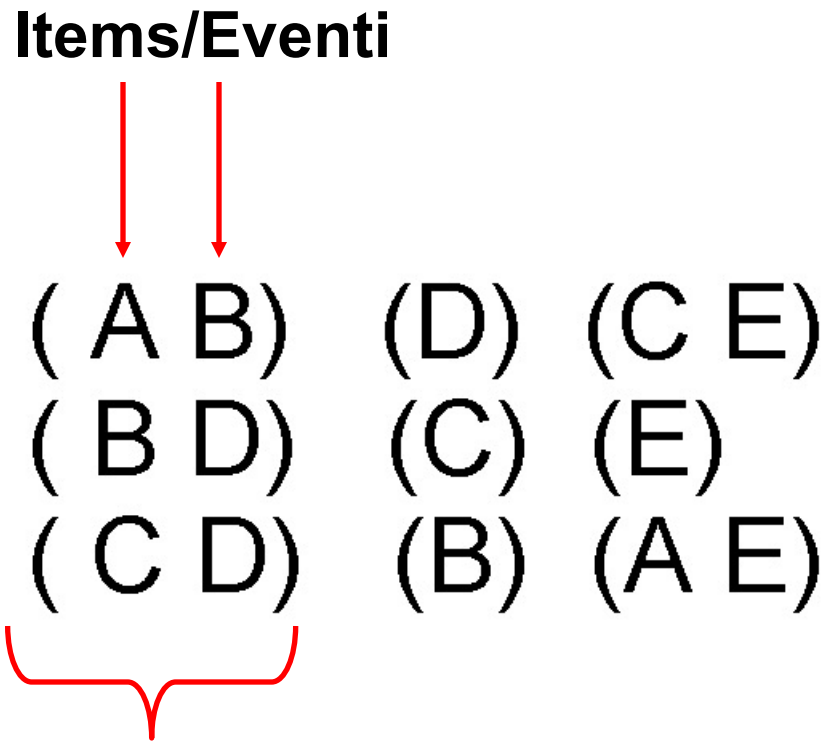
J. Ross Quinlan, "C4.5: Programs for Machine Learning", Morgan Kaufmann Publishers, 1993.
Michael Berry and Gordon Linoff, "Data Mining Techniques (For Marketing, Sales, and Customer Support), John Wiley & Sons, 1997.

General Data Mining

Usama Fayyad, "Mining Databases: Towards Algorithms for Knowledge Discovery", Bulletin of the IEEE Computer Society Technical Committee on data Engineering, vol. 21, no. 1, March 1998.

Christopher Matheus, Philip Chan, and Gregory Piatetsky-Shapiro, "Systems for knowledge Discovery in databases", IEEE Transactions on Knowledge and Data Engineering, 5(6):903-913, December 1993.

Dati sequenziali

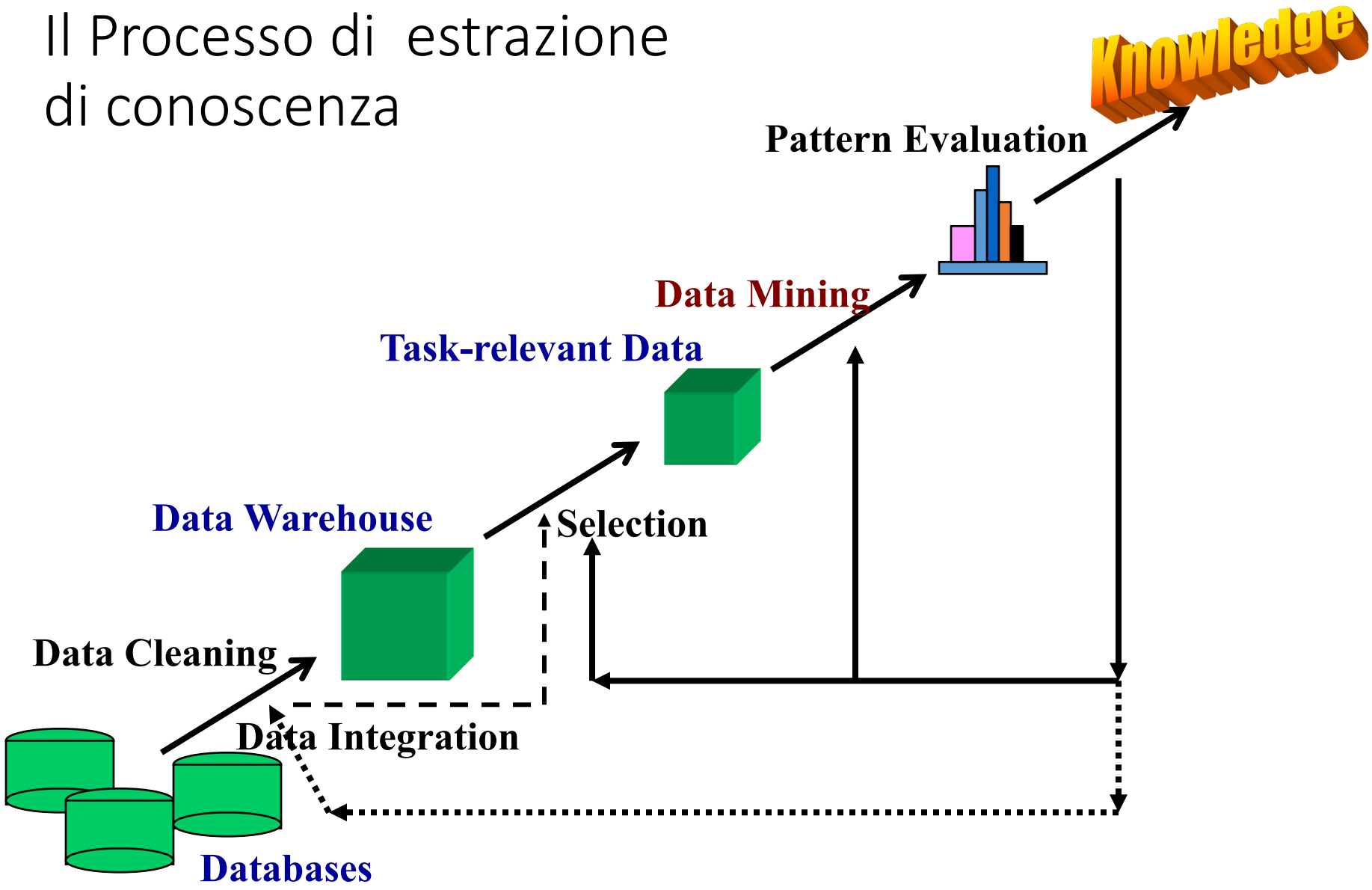


Dati sequenziali

- DNA

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

Il Processo di estrazione di conoscenza



Learning Supervisionato e non Supervisionato

- **Learning Supervisionato (classificazione)**
  - Supervisione: I dati usati per costruire il modello(osservazioni, misure, ecc.) sono associate con delle **etichette che indicano la classe** dell'oggetto osservato
  - **Sulla base del dato di training osservato si classificano i nuovi oggetti**
- **Learning non Supervisionato (clustering)**
  - Le etichette sul dato di training **non sono conosciute**
  - L'obiettivo è di trovare gruppi di oggetti simili senza l'uso di etichettature

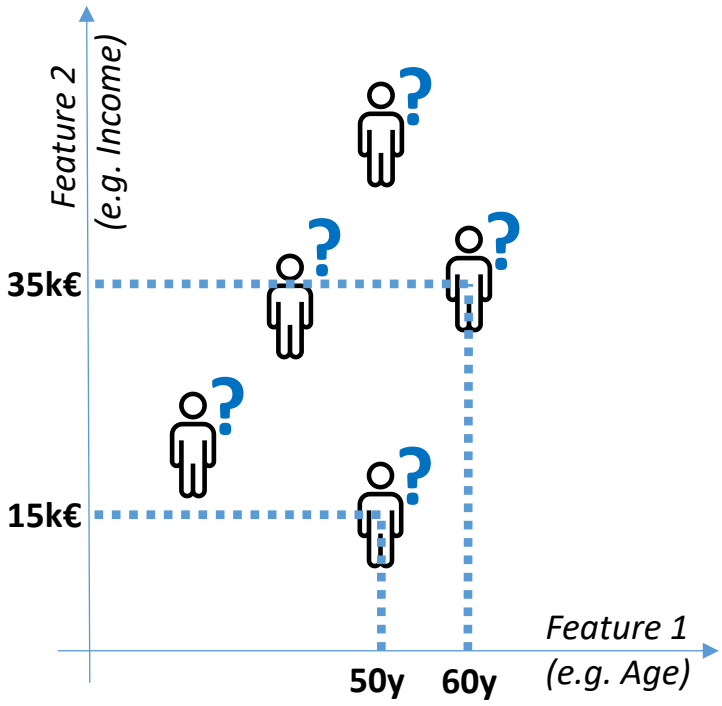
1

DATA MINING & MACHINE LEARNING

Classificazione e Predizione

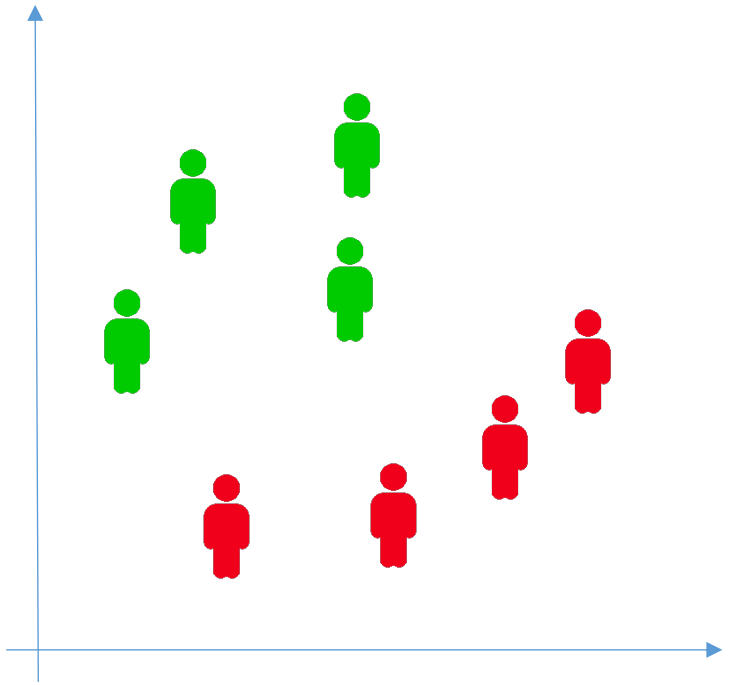
Il problema della classificazione

- **Cosa conosciamo**
  - Un insieme di **oggetti** descritti attraverso delle **caratteristiche**:
    - Persone descritte attraverso l'età, il sesso, l'altezza, ecc.
    - Transazioni bancarie descritte attraverso il tipo, la data e l'ammontare, ecc.
- **Cosa vogliamo fare**
  - Trovare un modello che impara a riconoscere la relazione tra alcuni attributi e un'etichetta (classe)
  - Associare gli oggetti a una classe, presa da una lista predefinita
    - "cliente fedele" 🟢 vs. "churner" 🔴
    - "transazione normale" 🟢 vs. "fraudolenta" 🔴
    - "Paziente a basso rischio" 🟢 vs. "alto rischio" 🔴



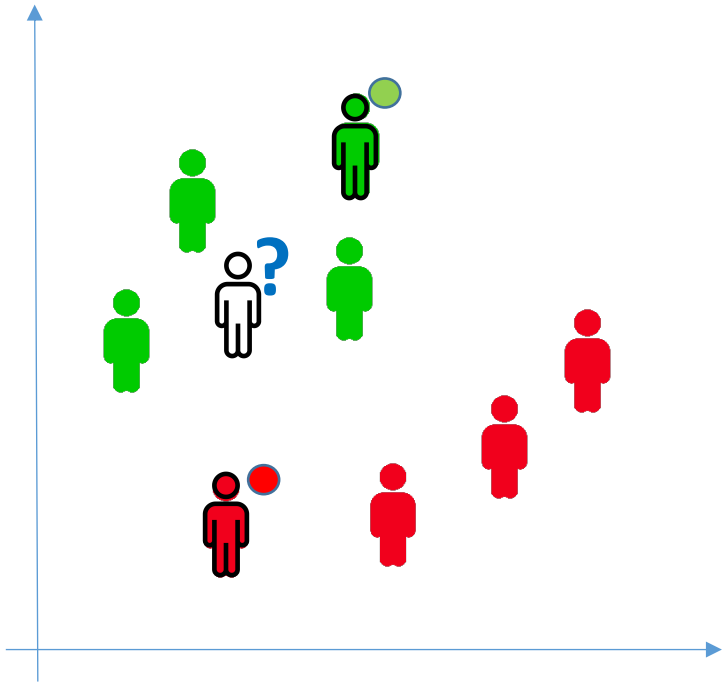
Il problema della classificazione

- Cosa conosciamo
  - Nessuna conoscenza di dominio
  - Solo esempi: Training Set
    - Oggetti etichettati
- Cosa possiamo fare
  - Imparare dagli esempi
  - Fare inferenze sugli altri oggetti



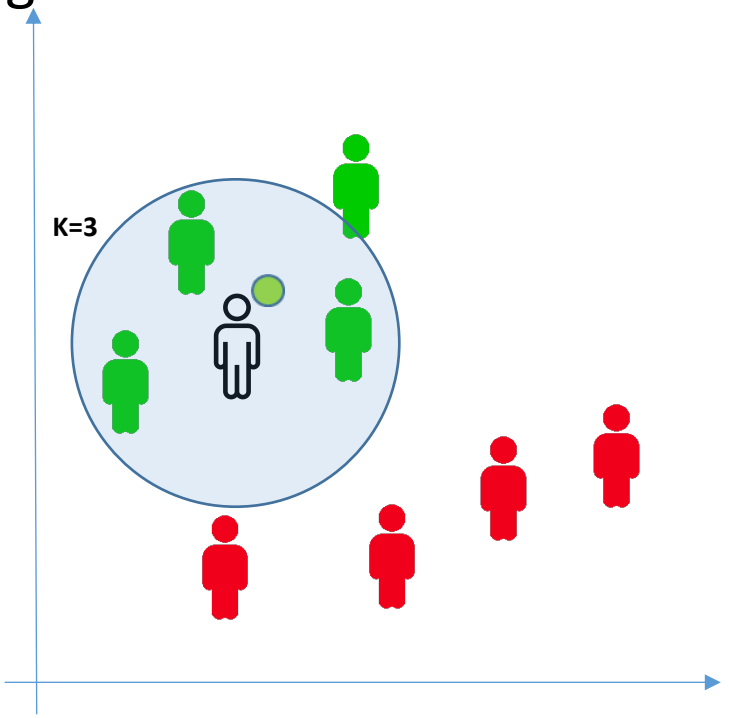
Il classificatore più banale

- Rote learner
  - Per classificare l'ogetto X, controlalre se nel training set esiste un esempio etichettato identico a X
  - Si → assegnare a X la stessa etichetta
  - No → Non so



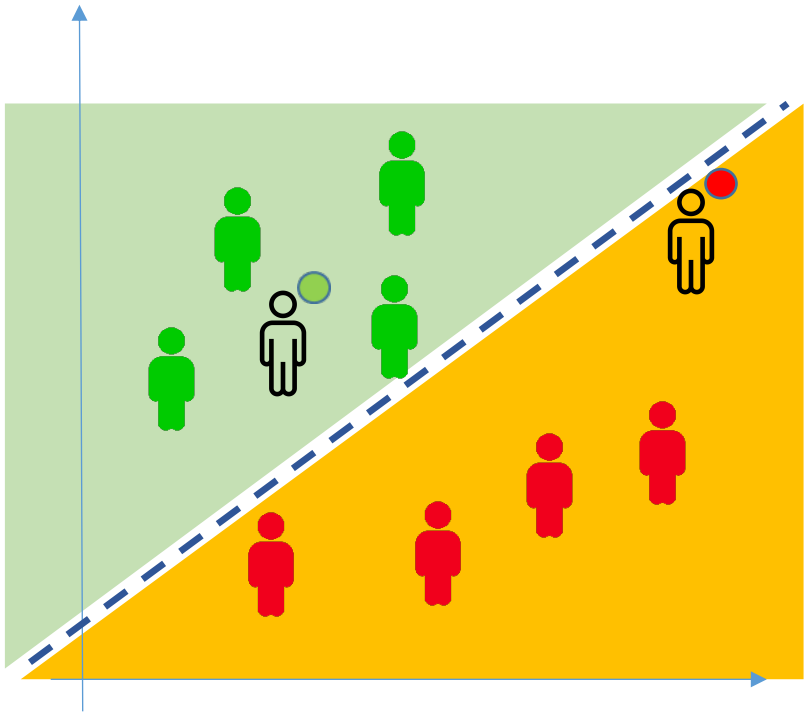
Classificare per similarità

- K-Nearest Neighbors
  - Assegnare l'etichetta dei K ogetti più simili che troviamo nel training set



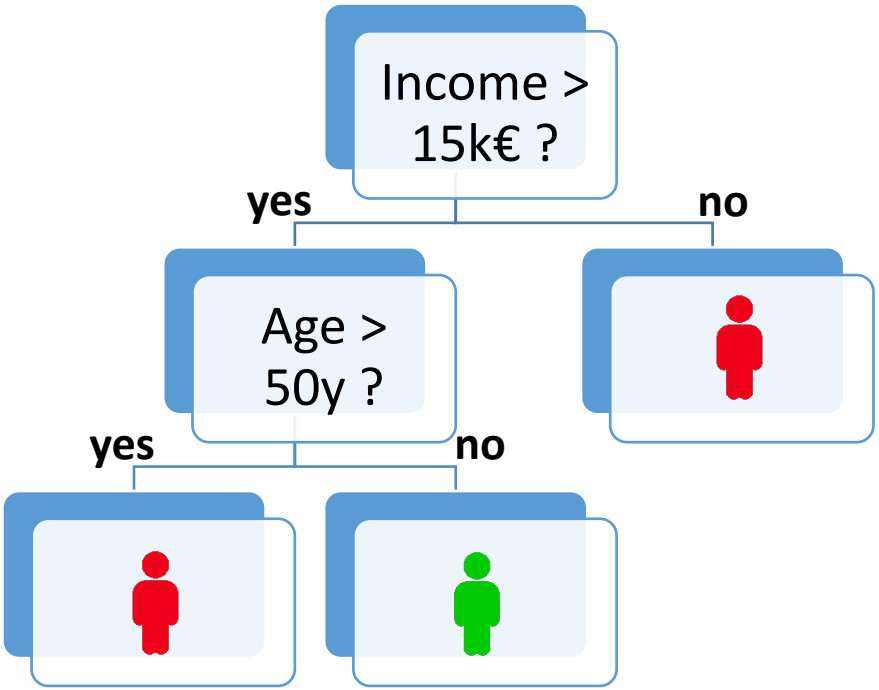
Costruire un modello

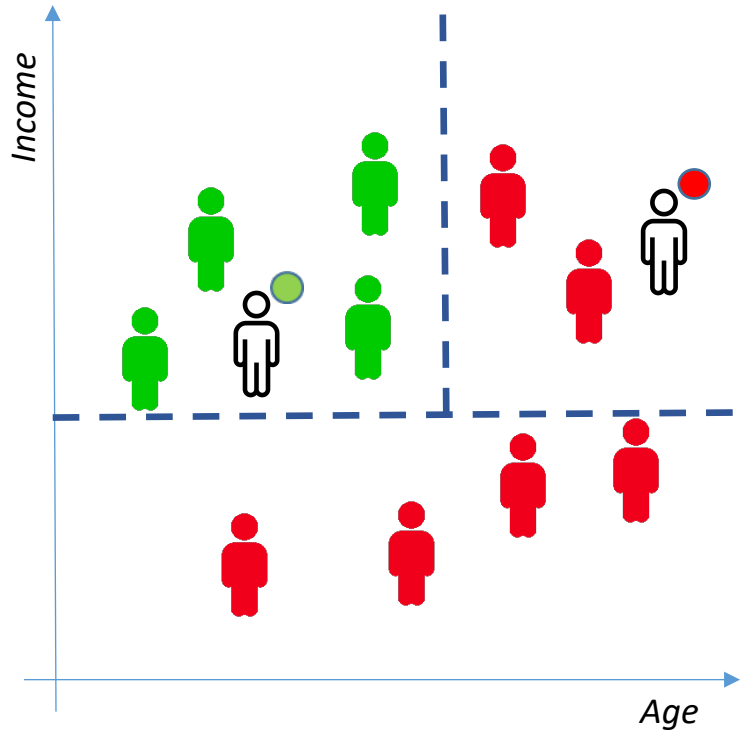
- Trovare una linea di separazione per dividere gli oggetti delle due classi



Costruire un modello

Decision Trees





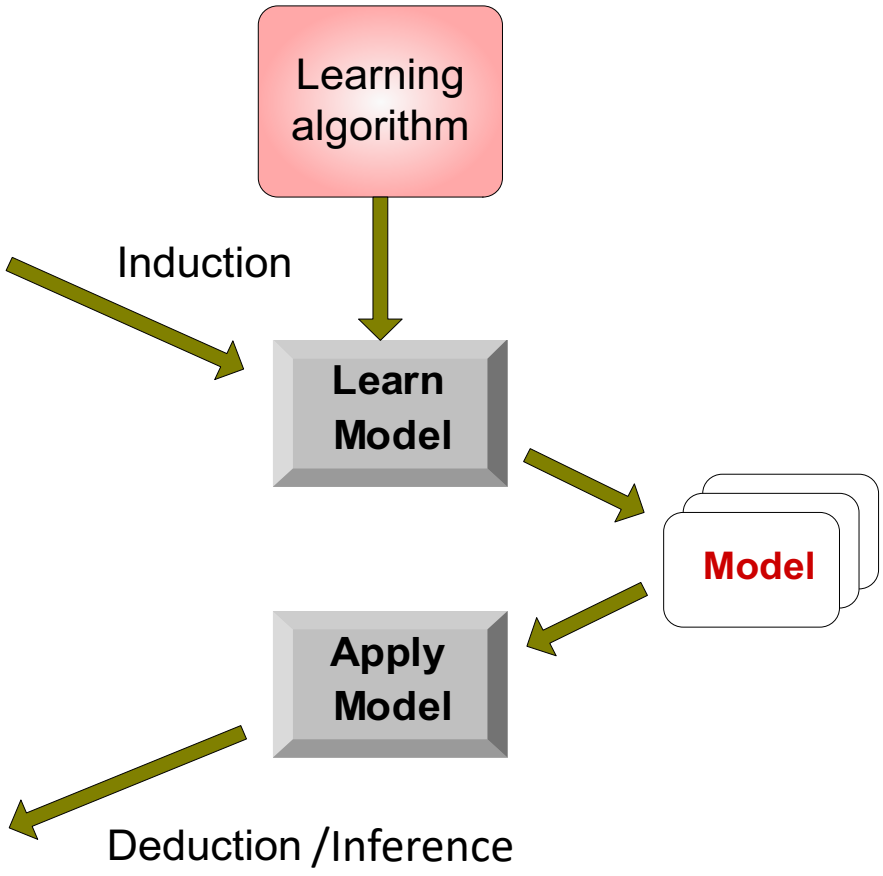
Approccio generale per l'apprendimento di un modello di classificazione

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



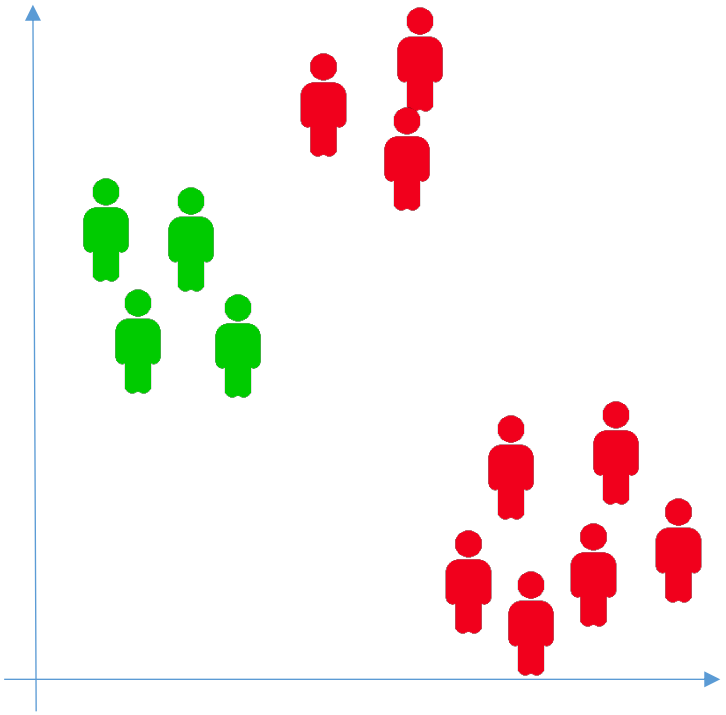
2

DATA MINING & MACHINE LEARNING

Clustering

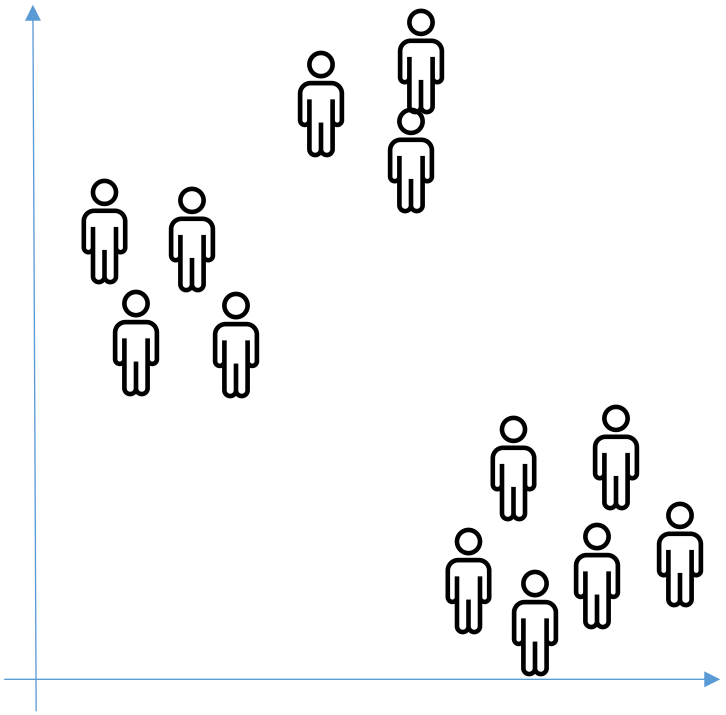
Clustering

- La classificazione sfrutta la conoscenza di oggetti per cui l'etichetta è conosciuta



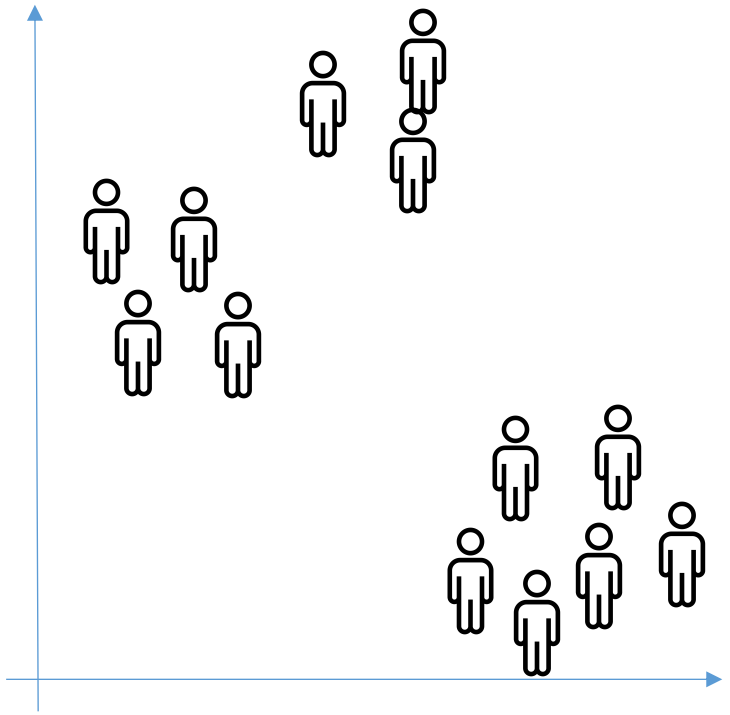
Clustering

- Cosa possiamo fare se le etichette non sono conosciute?
  - Potremmo avere le etichette solo per alcuni
  - Le etichetto potrebbero non esistere



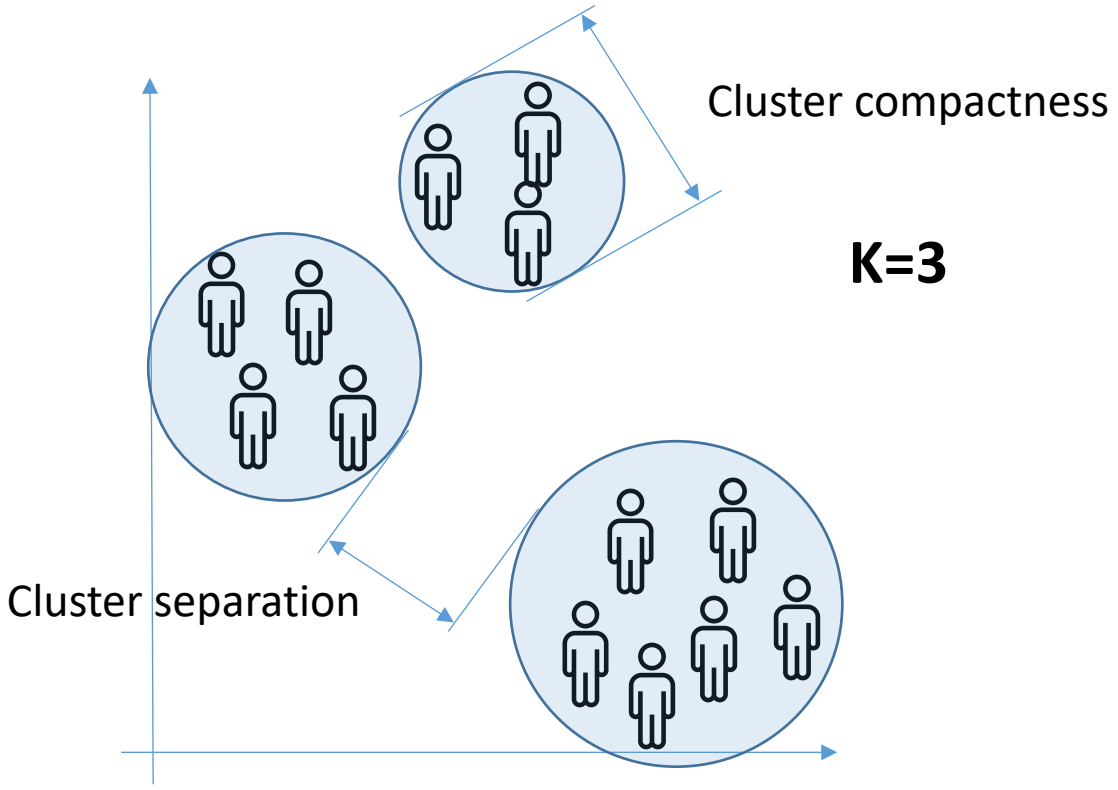
Clustering

- Goal: trovare una struttura nei dati
- Gruppi di oggetti che che sono simili sulla base di caratteristiche conosciute



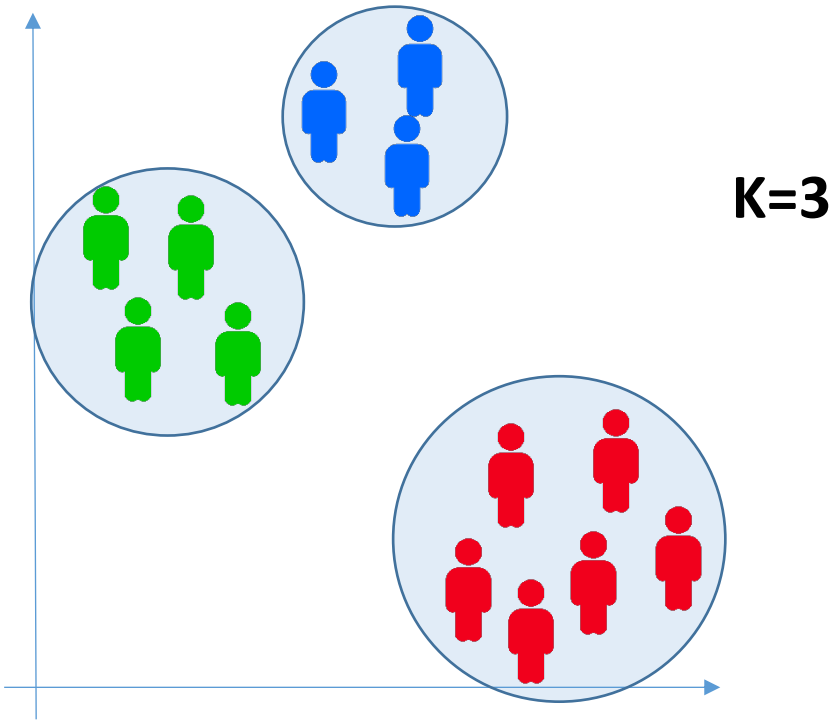
Clustering: K-means (family)

- Trova k sottogruppi che formano gruppi compatti



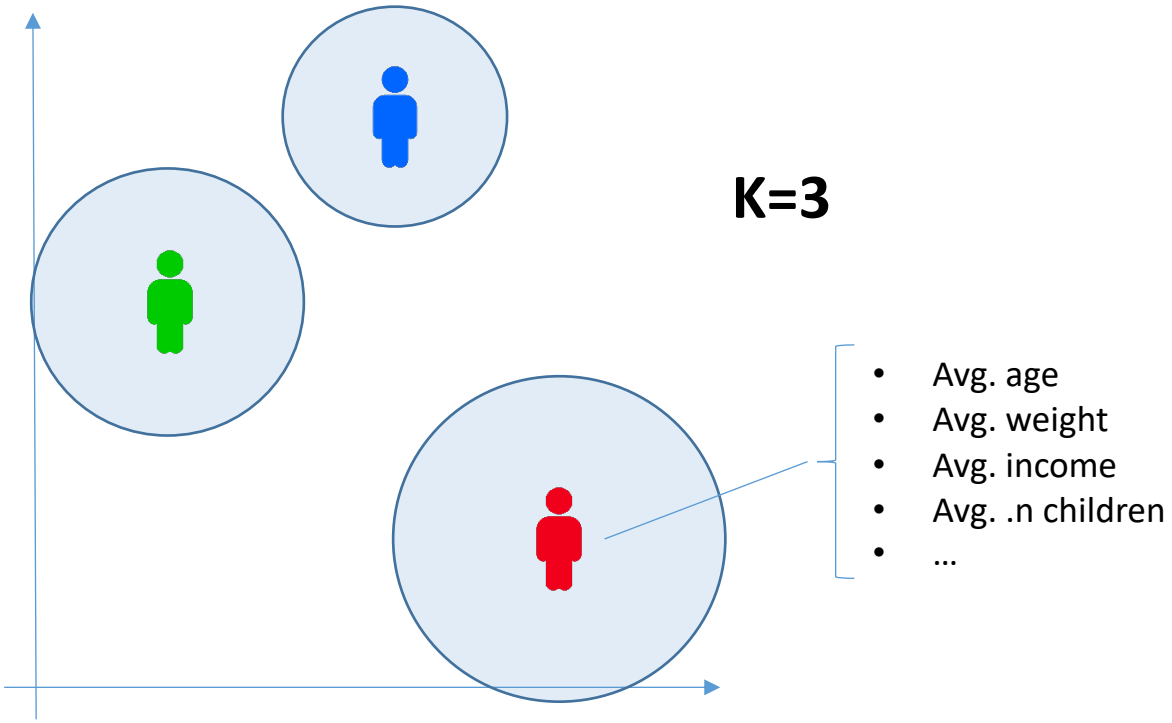
Clustering: K-means (family)

- Output 1: partizionare l'insieme degli oggetti in input



Clustering: K-means (family)

- Output 2: identificare K oggetti rappresentativi (centroidi)
- Centroide = profile medio degli oggetti del cluster



Customer Segmentation

Goal: Torvare una suddivisione dei **clienti** per indirizzare la campagna di marketing

Approccio

1. **Raccogliere informazioni diverse sui client:** info demografiche, stili di vita, comportamento di acquisto
2. **Trovare clusters** di client simili
3. Misurare la qualità dei **cluster** studiando se le caratteristiche dei clienti di uno stesso cluster sono simili

Risultato della segmentazione

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

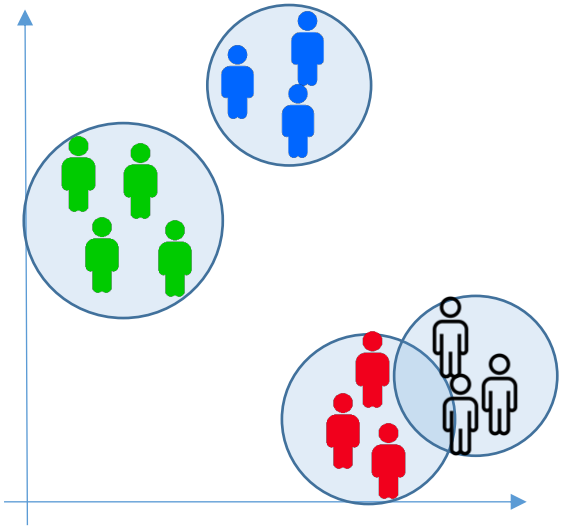
- Full store shop
- High avg. basket size / # trips

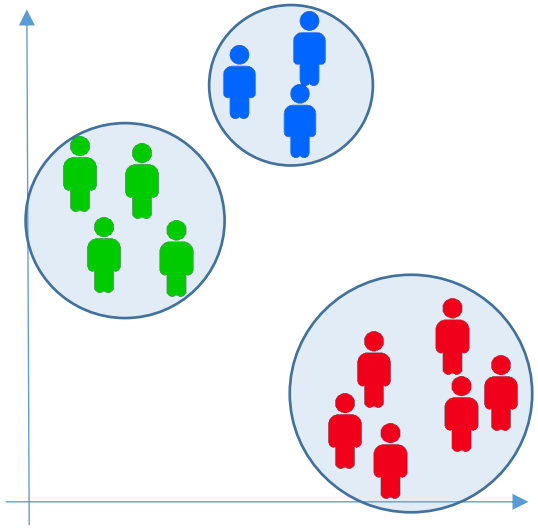
- High % purchased on promotion
- Rewards seeker

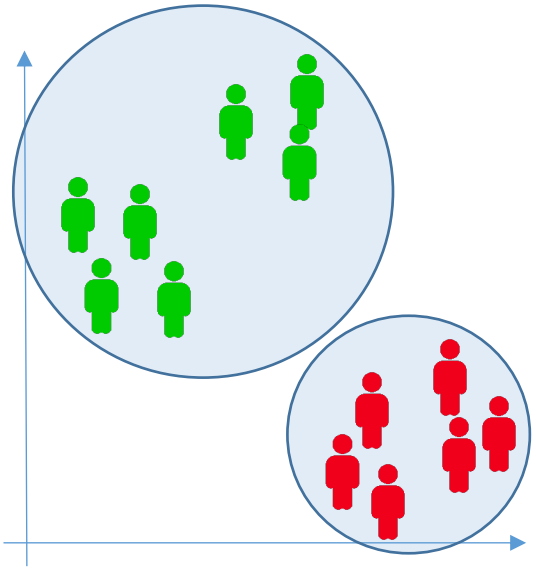
- High spend categories
  - Fresh produce
  - Organic food
  - Multipack juice, snack

Clustering: approccio gerarchico

- A volte sono preferibili livelli multipli di aggregazione

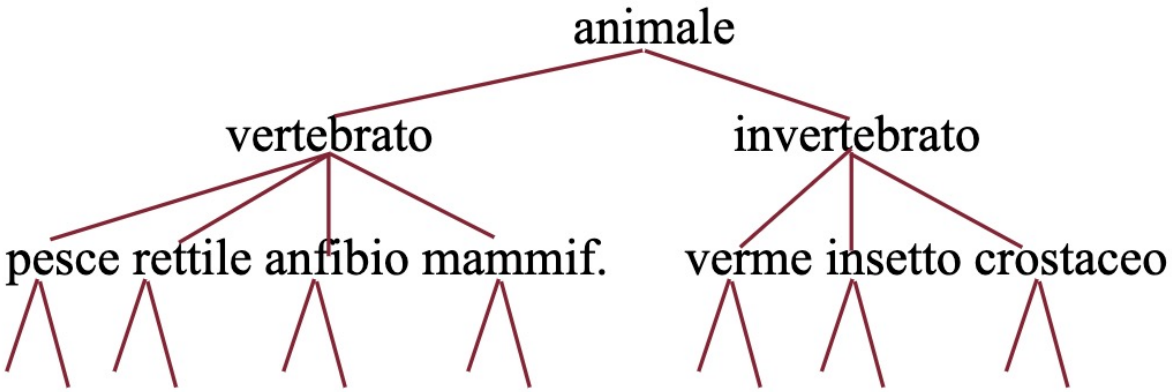




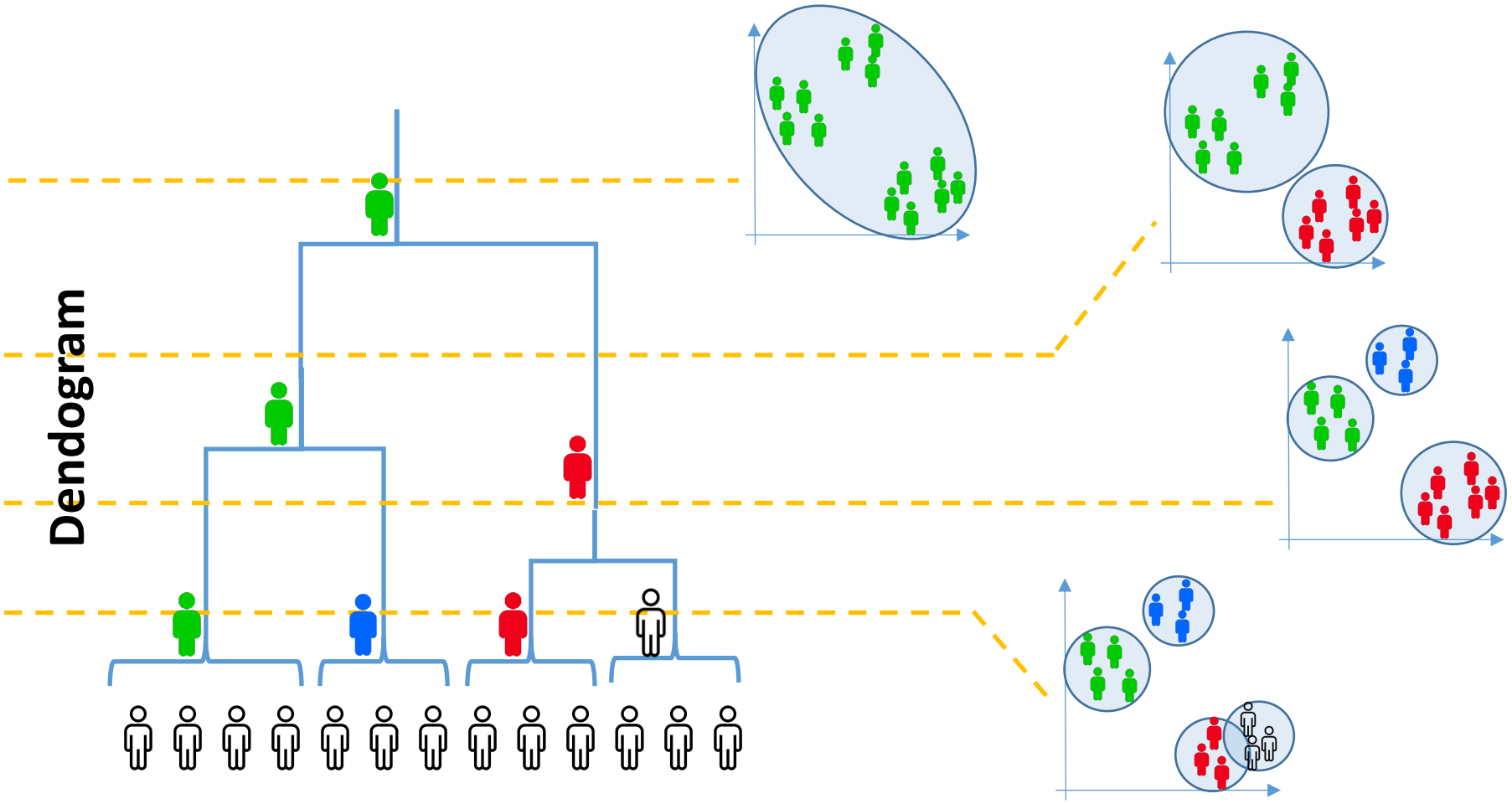


Clustering: approccio gerarchico

- Costruisce una tassonomia gerarchica ad albero a partire da oggetti non etichettati
- Assume una funzione di similarità per confrontare le caratteristiche degli oggetti



Clustering: approccio gerarchico



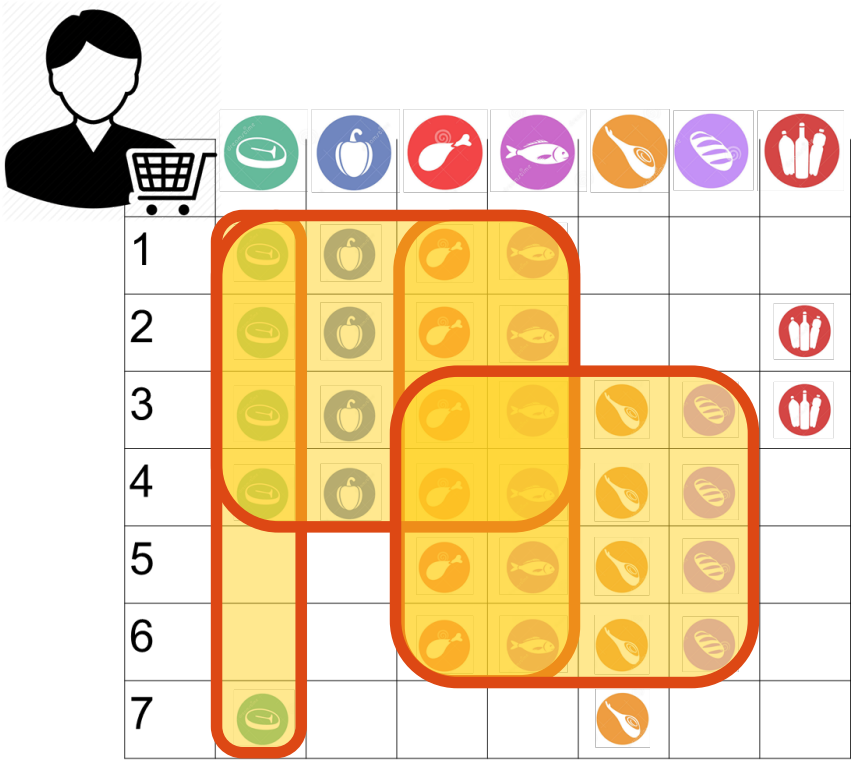
3

DATA MINING & MACHINE LEARNING

Frequent patterns

Pattern Frequenti

- Eventi che appaiono insieme nei dati
- Ad esempio prodotti acquistati dai clienti di un supermercato



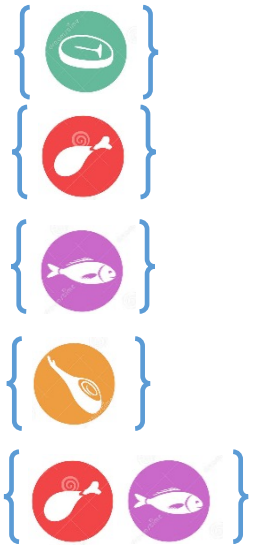
Estrarre le regole associative

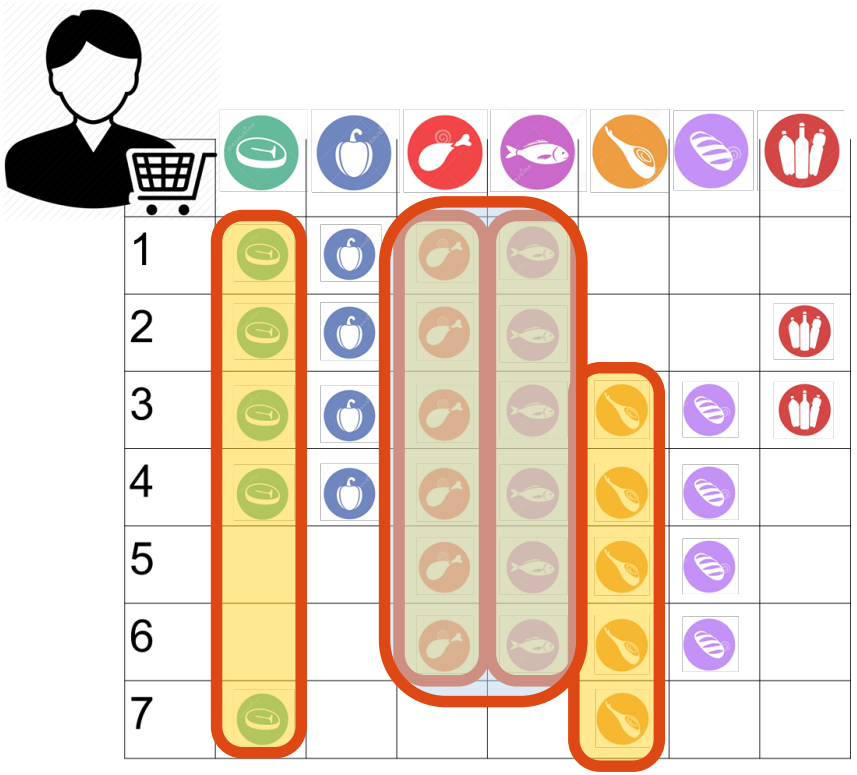
1. Generazione dei pattern frequenti
   - Generare tutti gli insiemi frequenti con frequenza $\geq$ minsup

2. Generazione Regole
   - Generare regole con alta confidenza a partire dai pattern frequenti

Pattern Frequenti

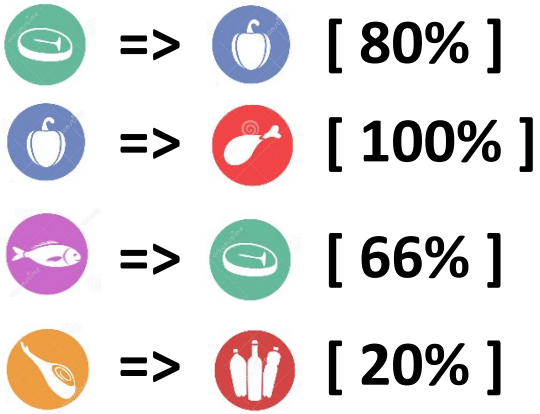
- **Frequent itemsets** rispetto a una soglia minima
- Per esempio con Min_freq = 5

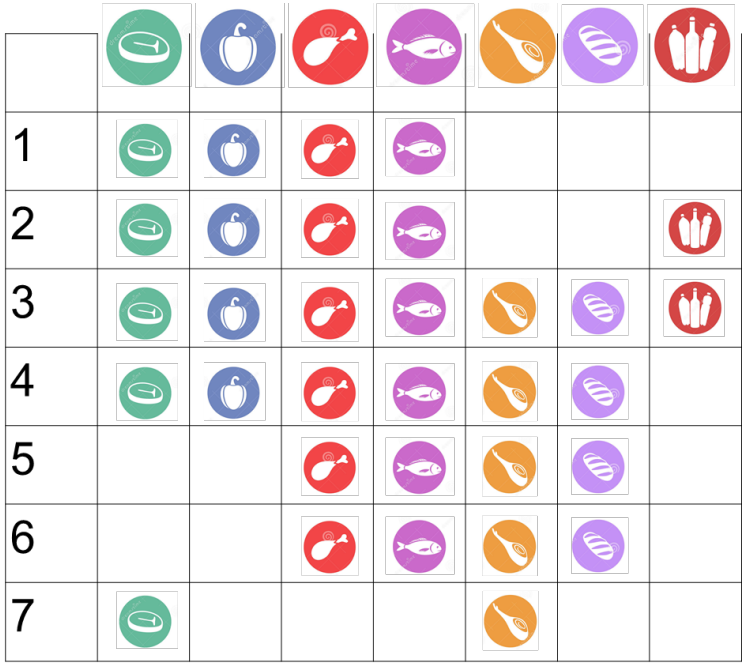




Regole Associative

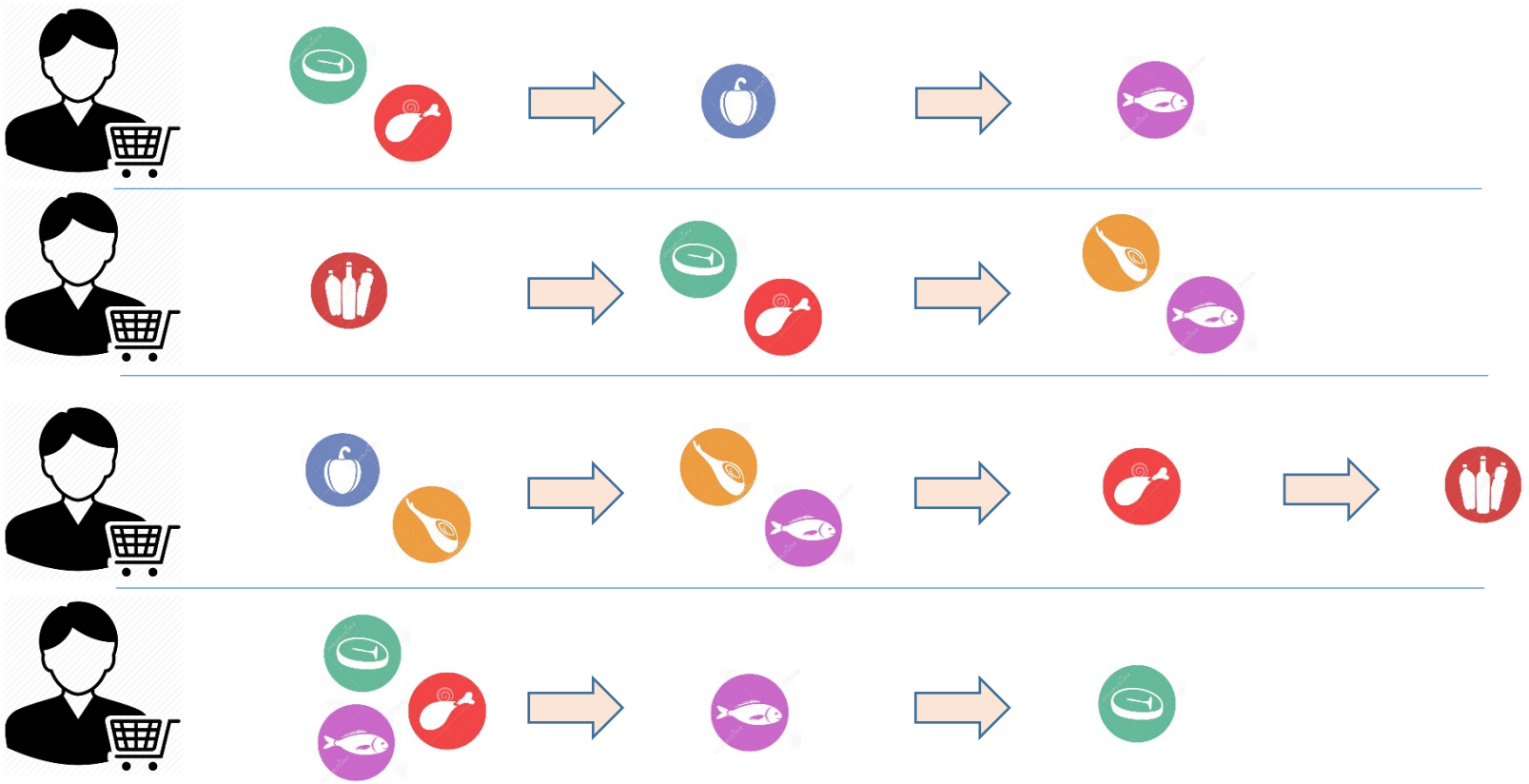
- Da insieme di eventi frequenti a regole if ... then
- Esempio: Se compriamo fragole allora è molto probabile che compreremo la panna -- probabilità dell'80% (confidenza)





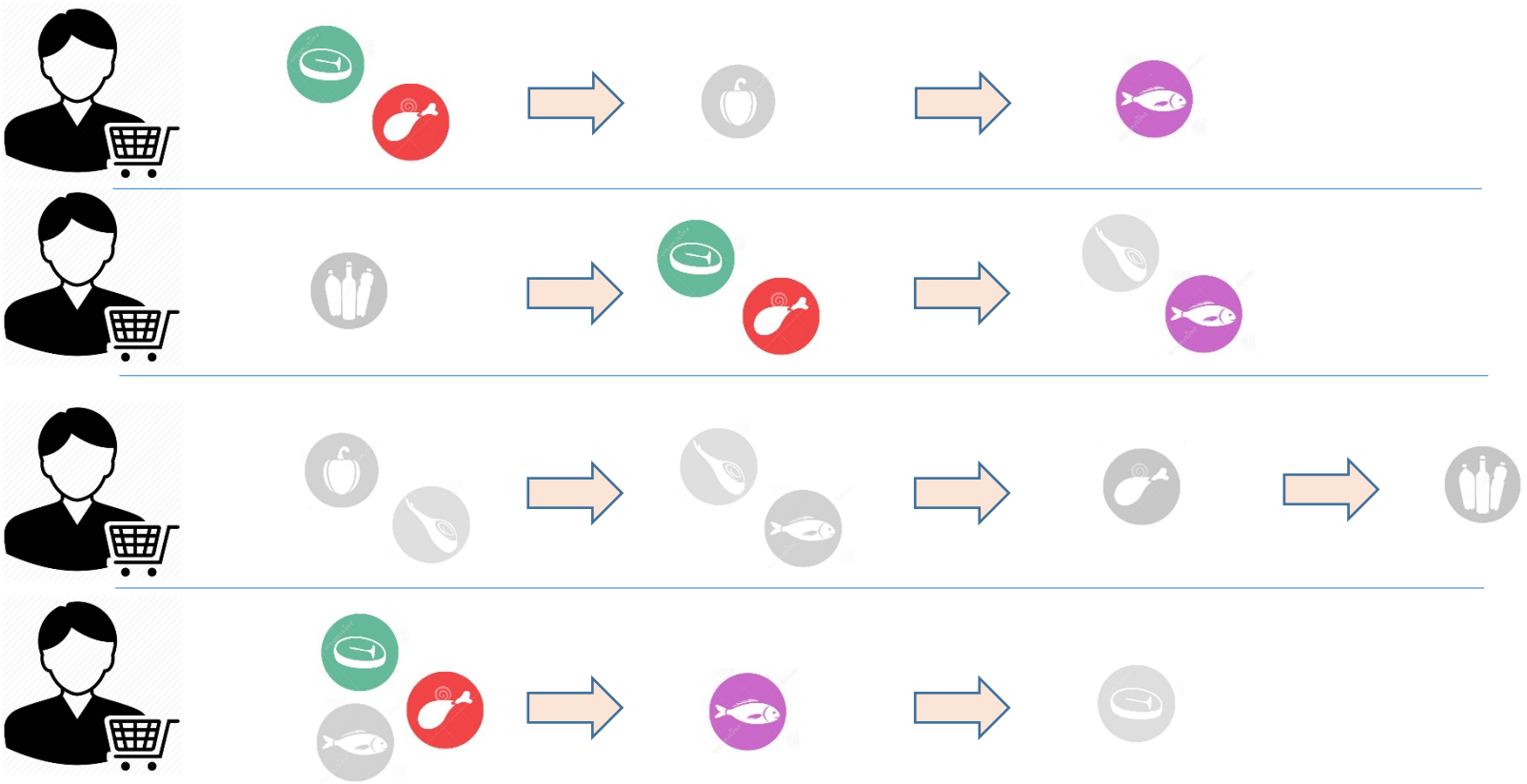
Pattern frequenti nelle sequenze

- Sequenze frequenti (a.k.a. pattern sequenziali)
- Input: sequenze di eventi (o di gruppi)



Pattern frequenti in sequenze

- Goal: identificare sequenze che occorrono frequentemente
- Pattern sequenziale: 



Breath & Summarize

Classification → Imparare da esempio per inferire la classe di nuovi oggetti
Clustering → Trovare una struttura nei dati
Frequent patterns → Trovare regolarità nei dati

THANK YOU !

Questions?