Exemplar Case Studies

Dino Pedreschi, Fosca Giannotti
Pisa KDD Lab, ISTI-CNR & Univ. Pisa
http://www-kdd.isti.cnr.it/

Master Mains 2018

Outline

Case studies from real data: Coop DW

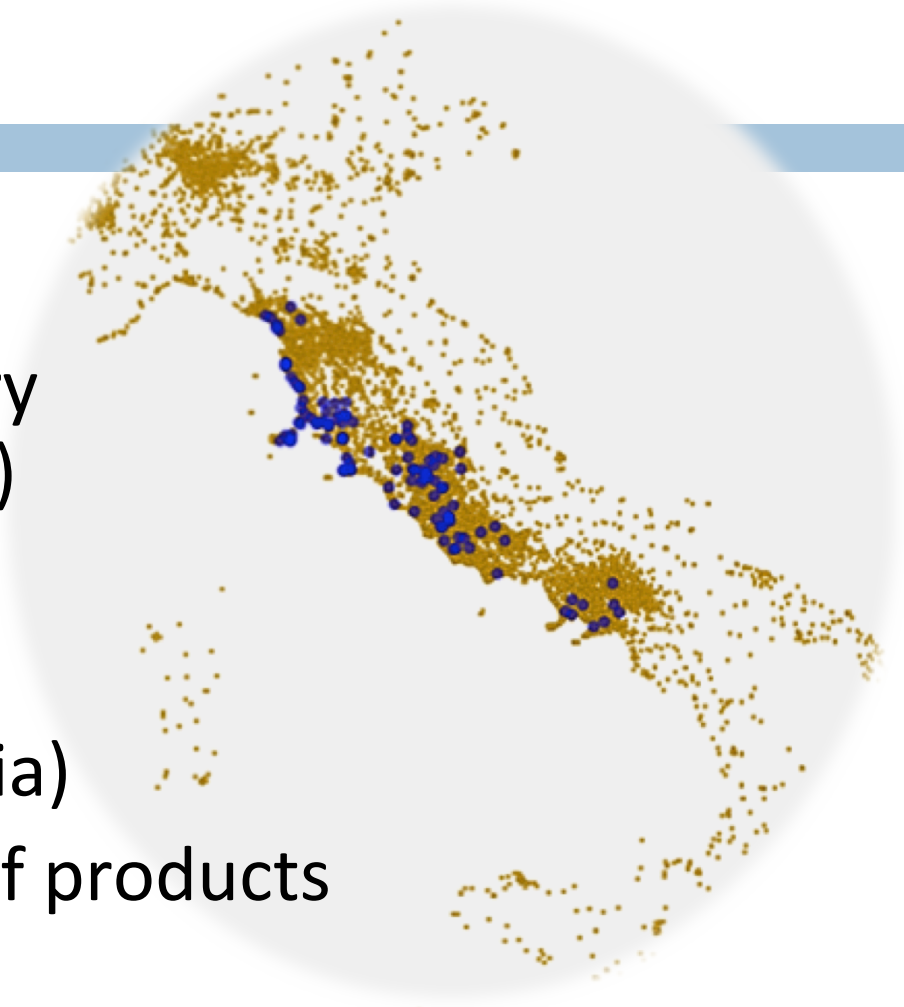
- CaseStudy1 - **Profitability & Predictability Indicators**
- CaseStudy2 – Personal Car Assistant
- CaseStudy3 – Sophistication Indicator & Nowcasting GDP

CaseStudy 4 – Now casting of flu trend with shopping behaviors

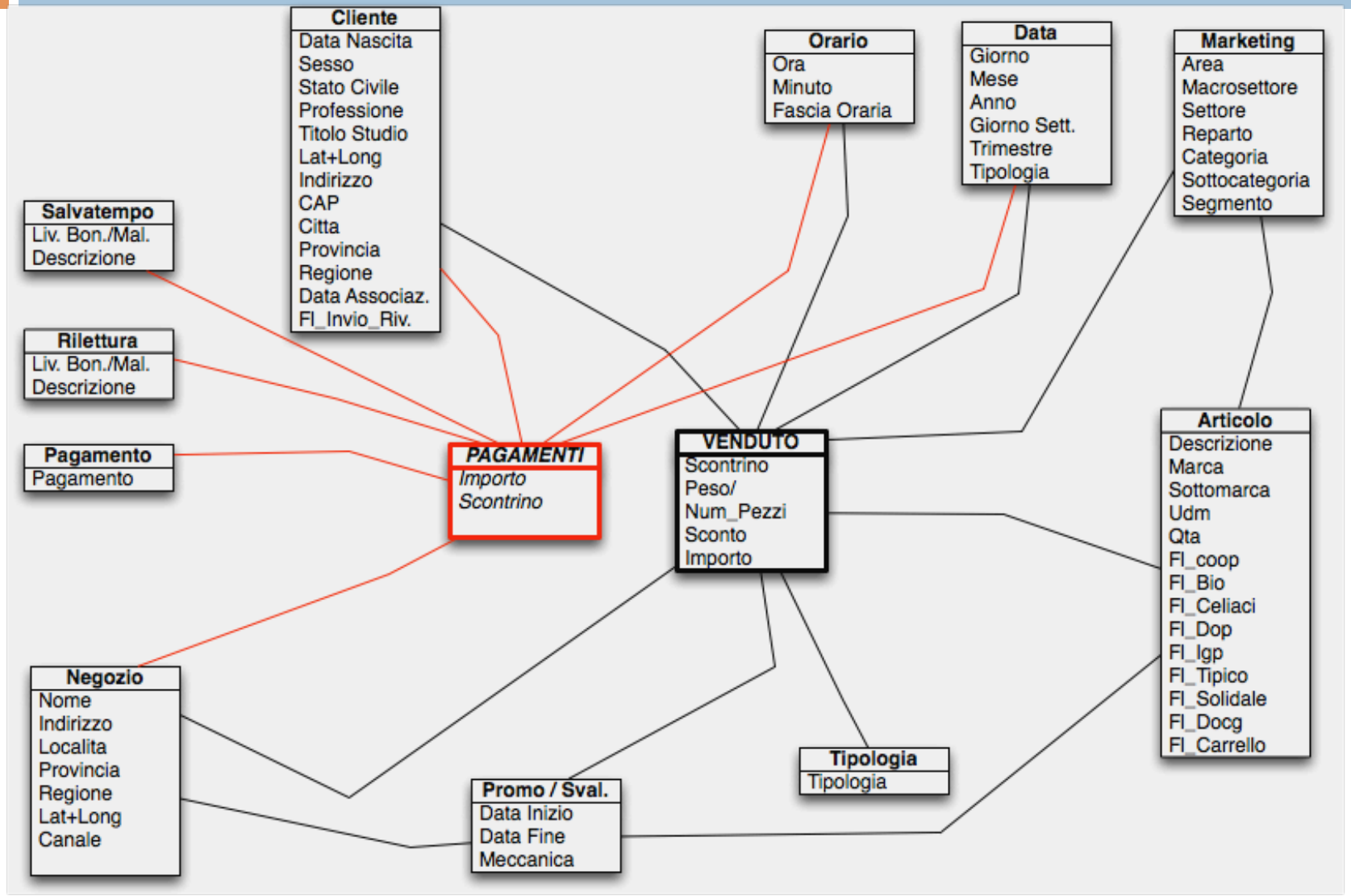
CaseStudy5 – Discovering Innovators

COOP Data

- More than 7 years of purchases (from January 2007 until August 2014)
- ~1 M active clients
- 138 shops (Tuscany, Umbria, Lazio, Campania)
- ~450K different types of products
- ~280M receipts
- ~280G product scans



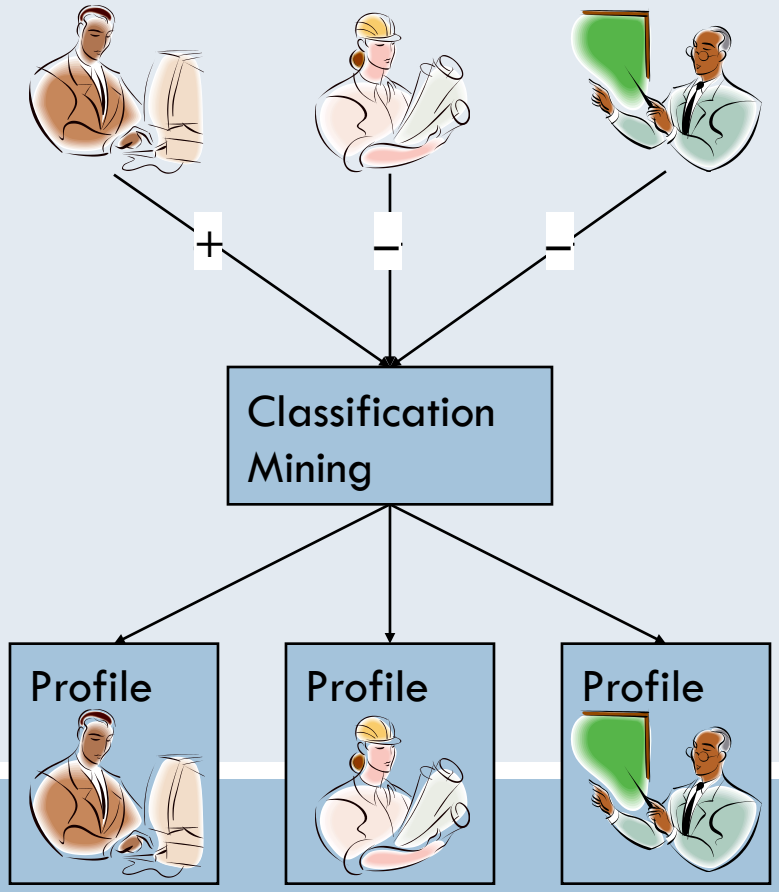
COOP Data


Cliente
Data Nascita
Sesso
Stato Civile
Professione
Titolo Studio
Lat+Long
Indirizzo
CAP
Citta
Provincia
Regione
Data Associaz.
Fl_Invio_Riv.
Orario
Ora
Minuto
Fascia Oraria
Data
Giorno
Mese
Anno
Giorno Sett.
Trimestre
Tipologia
Marketing
Area
Macrosettore
Settore
Reparto
Categoria
Sottocategoria
Segmento
Salvatempo
Liv. Bon./Mal.
Descrizione
Rilettura
Liv. Bon./Mal.
Descrizione
Pagamento
Pagamento
PAGAMENTI
Importo
Scontrino
VENDUTO
Scontrino
Peso/
Num_Pezzi
Sconto
Importo
Articolo
Descrizione
Marca
Sottomarca
Udm
Qta
Fl_coop
Fl_Bio
Fl_Celiaci
Fl_Dop
Fl_Igp
Fl_Tipico
Fl_Solidale
Fl_Docg
Fl_Carrello
Negozio
Nome
Indirizzo
Localita
Provincia
Regione
Lat+Long
Canale
Promo / Sval.
Data Inizio
Data Fine
Meccanica
Tipologia
Tipologia

INDVIDUAL PURCHAISING PROFILES

Obiettivo di analisi: realizzare un modello di segmentazione dei clienti basato sulla sistematicità degli acquisti. Tale sistematicità viene calcolata utilizzando due componenti: livello di regolarità dei prodotti nel carrello e livello di regolarità della componente spazio-temporale.

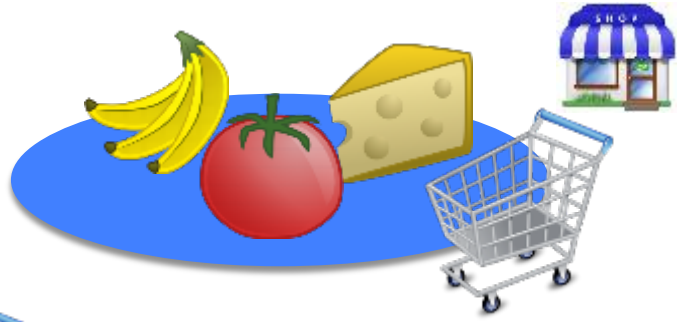
Profiling



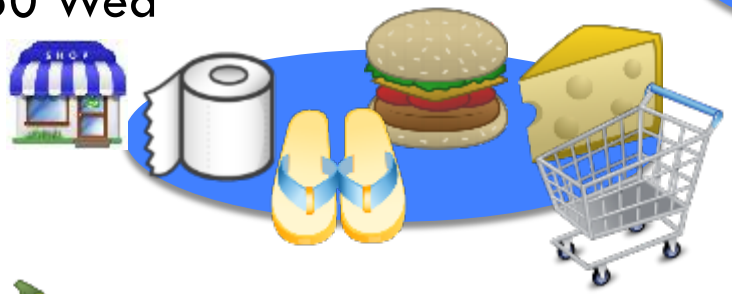
Personal Perspective

19.30 Sun

17.30 Mon



17.50 Wed



08.40 Mon

11.50 Tue

Extract Personal Representative Baskets

Predictability Indicators

Nuove Domanda: ci sono dei clienti sistematici? Quanto valgono? Rappresentano una categorie ? profiqua?(patter mining al lavoro!)

Data Preparation

- Filtro sulla dimensionalità dei dati
  - Si considera il solo venduto nell'anno 2012
  - Si considerano i soli negozi della provincia di Livorno
  - Si considerano solo i clienti frequenti (che abbiano fatto almeno una spesa al mese)

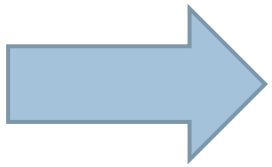
Data Preparation (cnt)

- Dimensioni del dataset dopo il filtro
  - 71.172.672 scansioni
  - 56.448 clienti
  - 84.362 articoli distinti
  - 23 negozi (1 Iper, 9 Super e 13 GestIn)

Data Preparation (cnt)

- Per il calcolo del BRI (indice di regolarità del carrello) bisogna applicare aPriori, quindi i dati vanno trasformati, per ogni cliente, da formato relazionale a formato transazionale

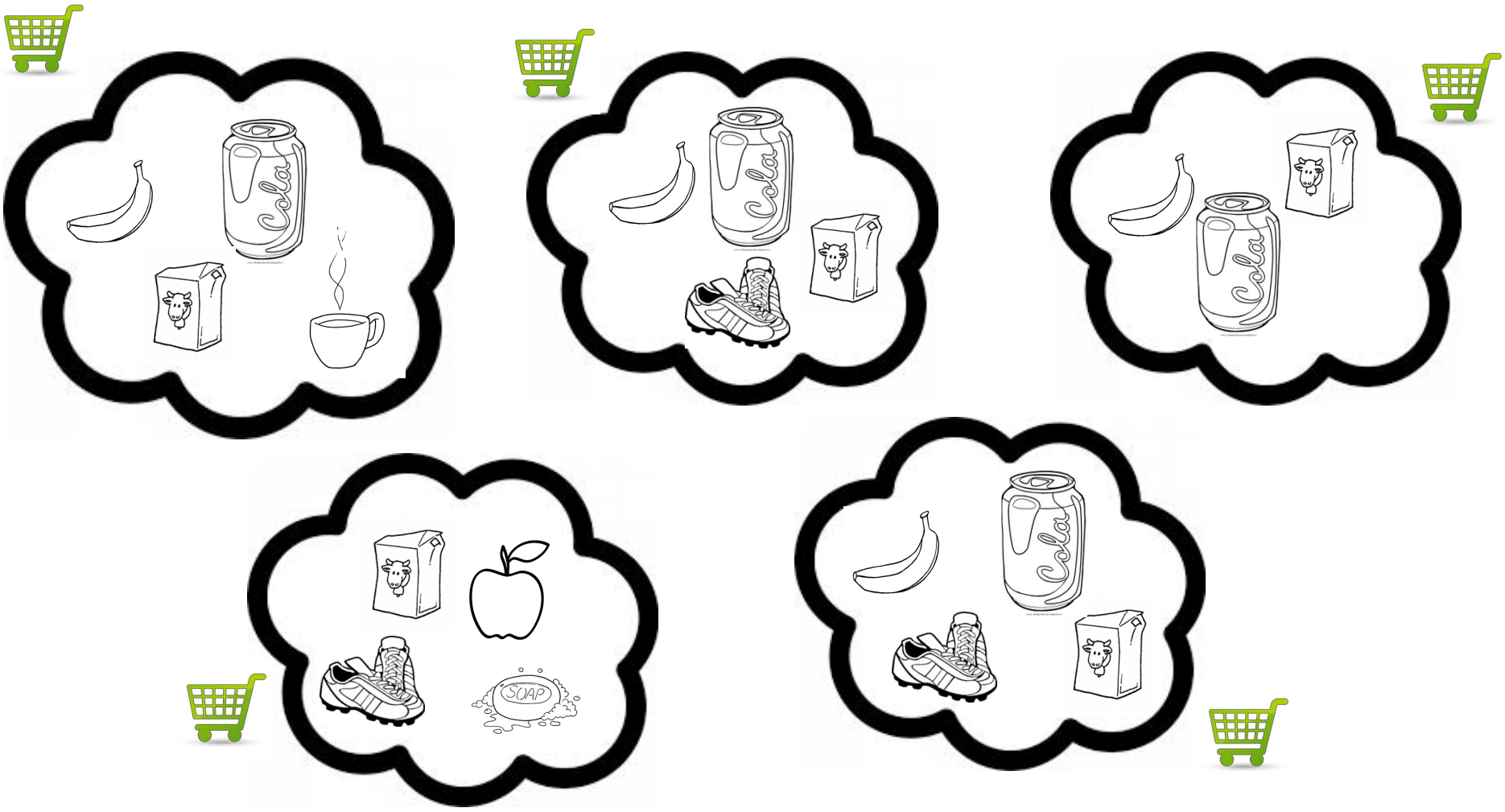
Scontrino	Prodotto
1	A
1	B
1	C
1	D
2	A
2	D
2	F



Scontrino	Lista Prodotti
1	A B C D
2	A D F

II Basket Regularity Index (BRI)

Starting from a set of baskets



II Basket Regularity Index (cnt)

Performing Frequent Pattern Mining (min supp=3)


Sup=5


Sup=4


Sup=4


Sup=4


Sup=4


Sup=3

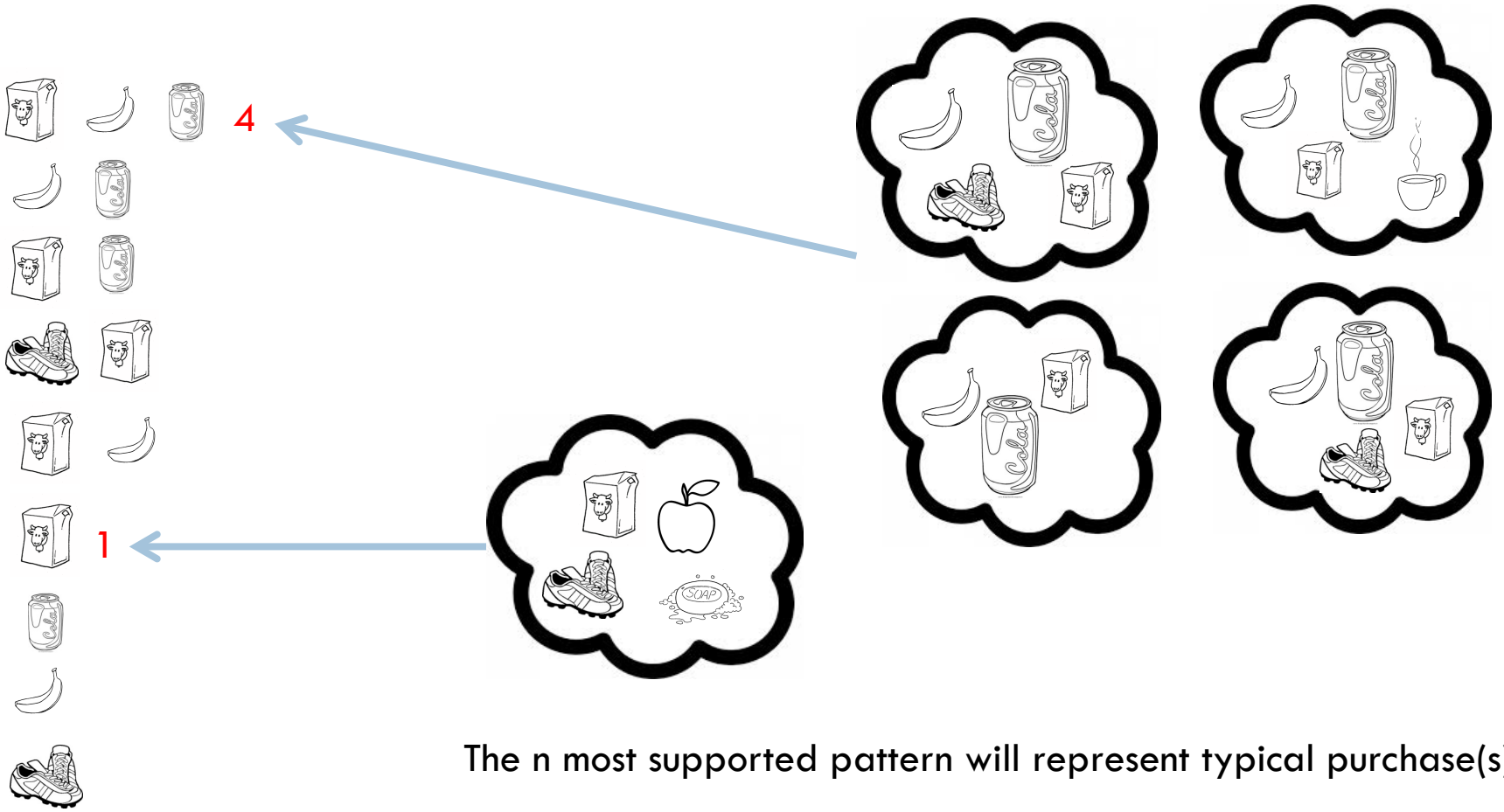

Sup=3


Sup=4


Sup=4

II Basket Regularity Index (BRI)

Each basket will support the longest pattern contained



The n most supported pattern will represent typical purchase(s).

II Basket Regularity Index (BRI)

- the Basket Regularity Index (BRI) a measure describing of the customer: it is the information Entropy calculated among all the supported patterns


4


1

$$BRI = -\left(\frac{4}{5}\log\frac{4}{5} + \frac{1}{5}\log\frac{1}{5}\right) = -\left(0.8 \cdot -0.321928 + 0.2 \cdot -2.321928\right) = 0.721928$$

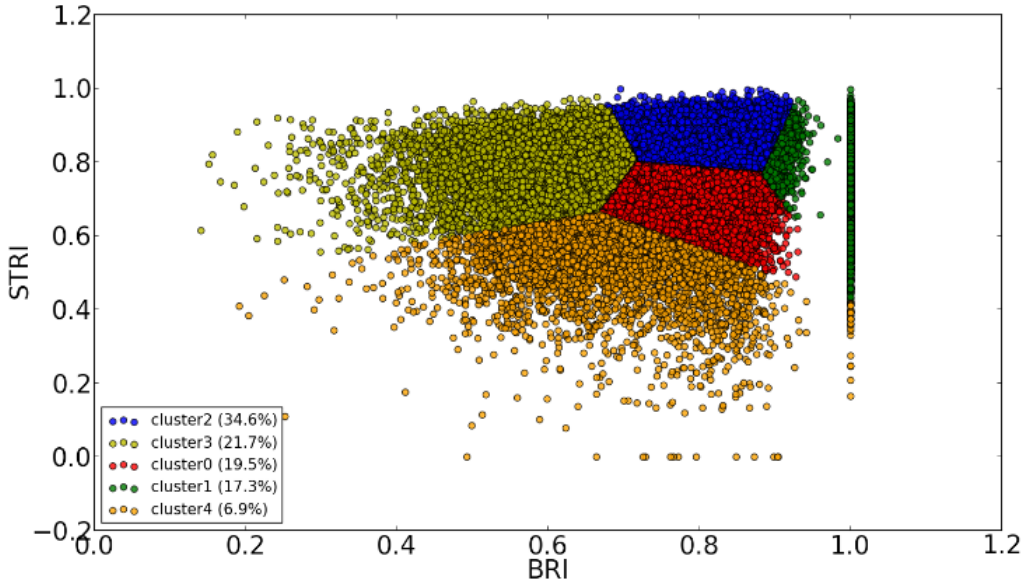
Lo Spatio-Temporal Regularity Index (STRI)

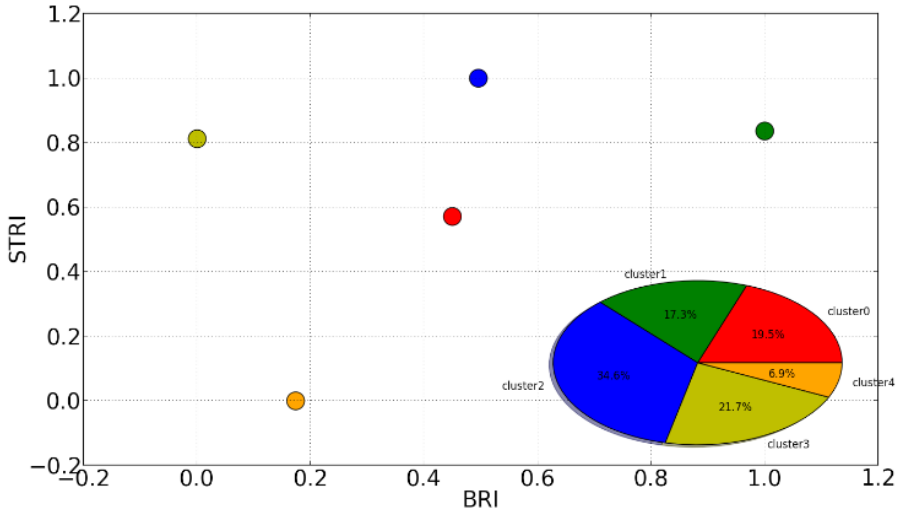
- Si salta il passaggio con aPriori, in quanto le informazioni sono più concise
- Si costruiscono triple contenenti informazioni su:
  - Negozio dove si è fatta la spesa
  - Tipo di giorno in cui si è fatta la spesa (feriale-festivo)
  - Fascia Oraria in cui si è fatta la spesa (mattina presto, tarda mattinata, primo pomeriggio, secondo pomeriggio, sera)
- Per ogni cliente, si calcola l'entropia su tali triple

Analisi: segmentazione

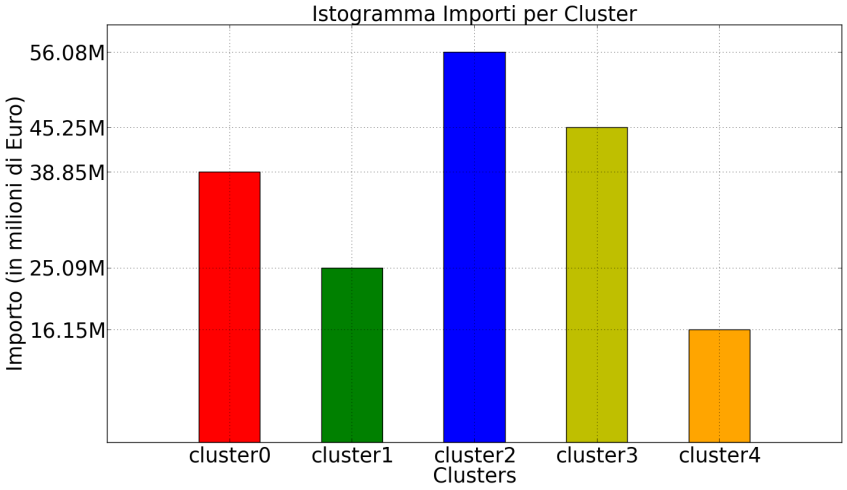
- Ad ogni cliente, quindi, sarà associata una coppia di indici (BRI e STRI)
- Queste coppie di indici possono essere riportate su un piano e sulle stesse possono essere applicati degli algoritmi di clustering
- In questo modo, la clientela verrà segmentata in base alla loro propensione ad essere sistematici per una delle due componenti principali del processo di spesa (cosa compro e quando/dove compro)

Analisi: segmentazione (cnt)

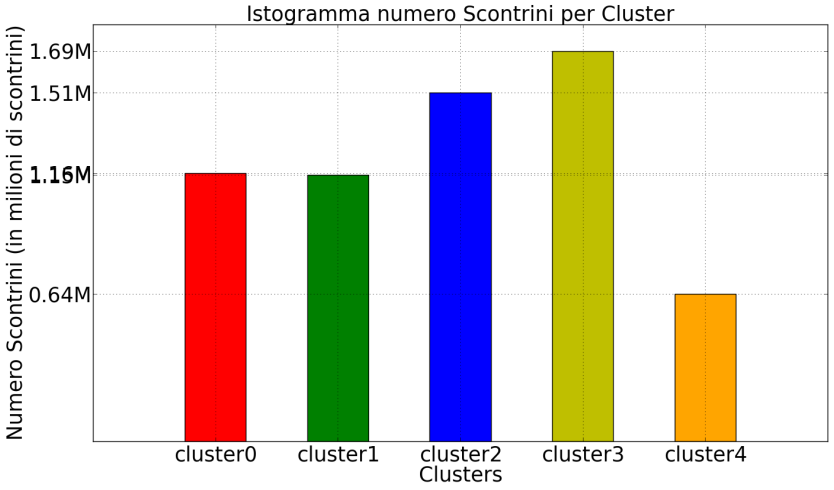




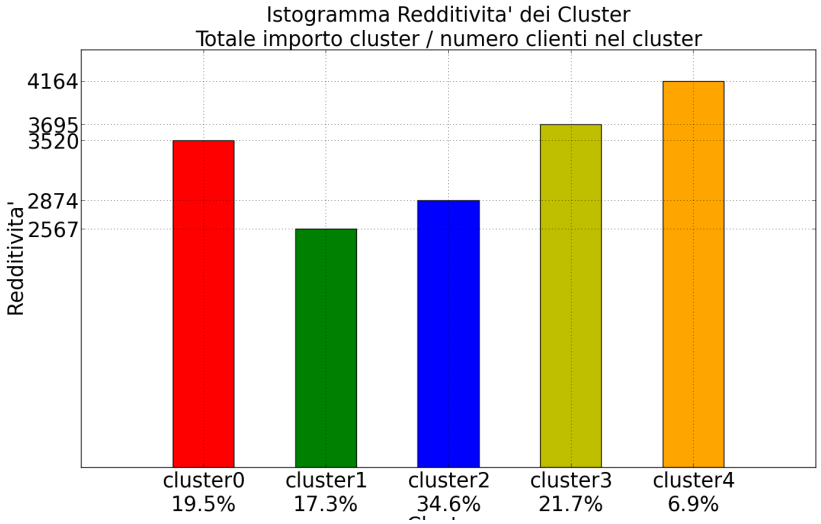
Analisi: segmentazione (cnt)



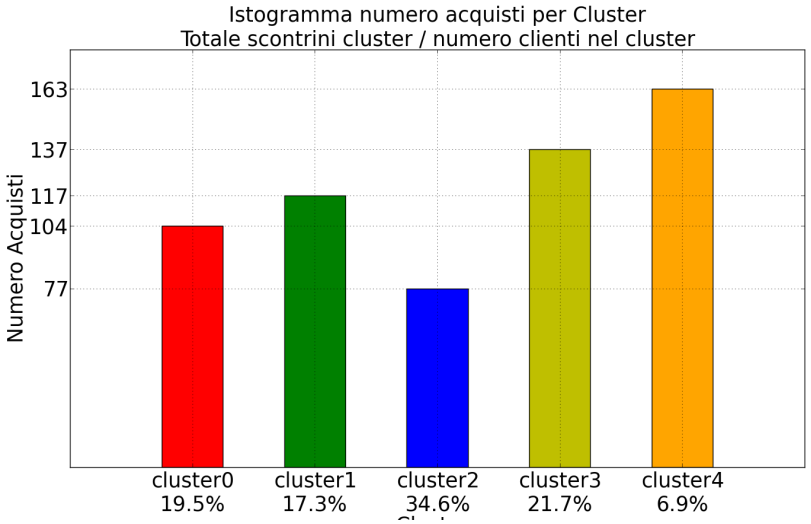
Tot importi



Tot scontrini



Importi medi

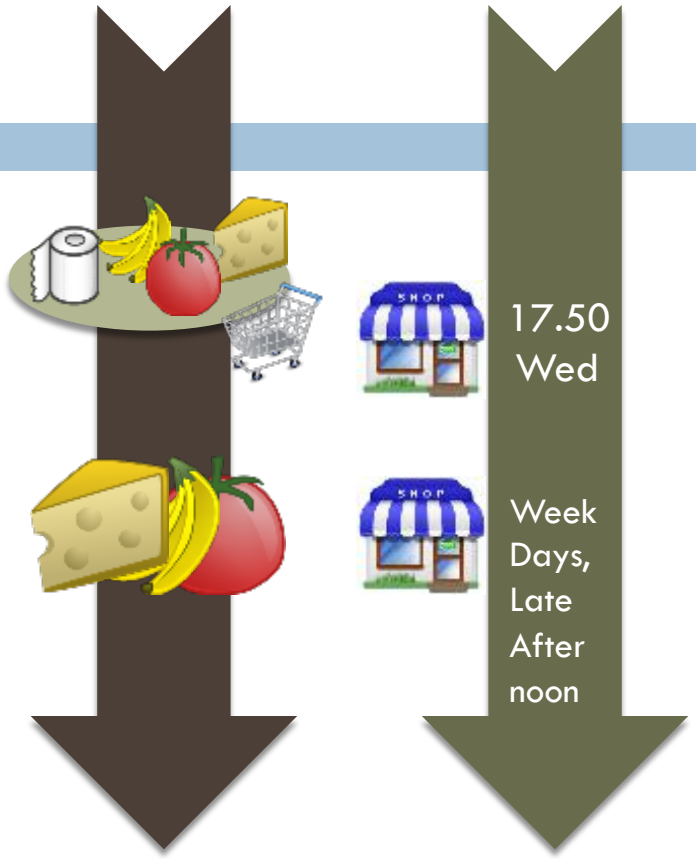


Scontrini medi

BRE & STRE Calculus

BRE: Basket Revealed Entropy
STRE: Spatio-Temporal Revealed Entropy

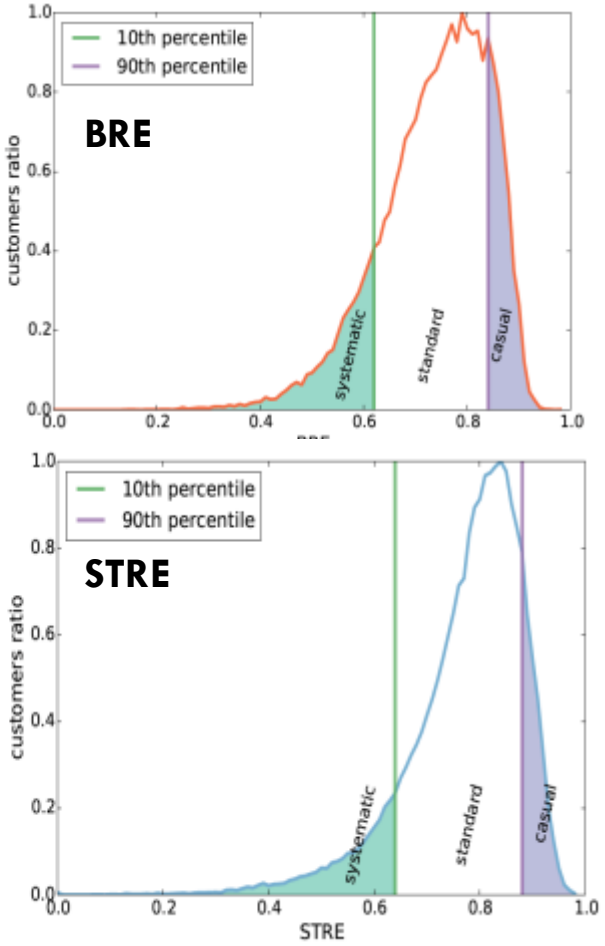
- ***Step 1:*** identify *representative baskets patterns* w.r.t. basket composition - shop and time.
- ***Step 2:*** classify each basket with a representative basket patter.
- ***Step 3:*** calculate BRE and STRE by using the frequencies and the entropy formula

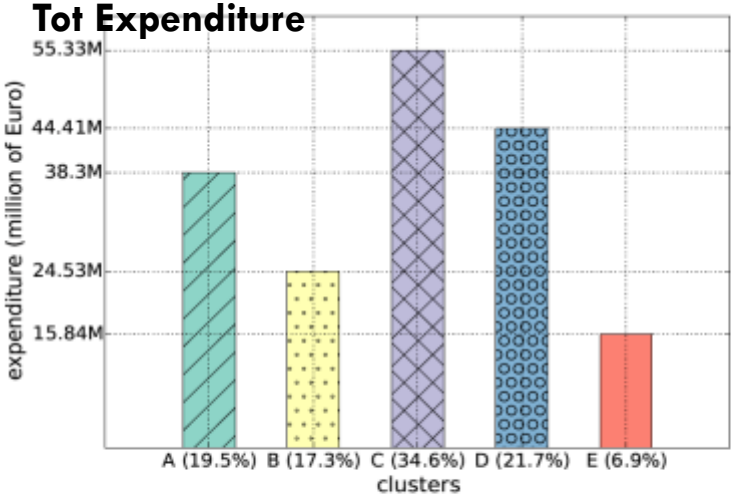


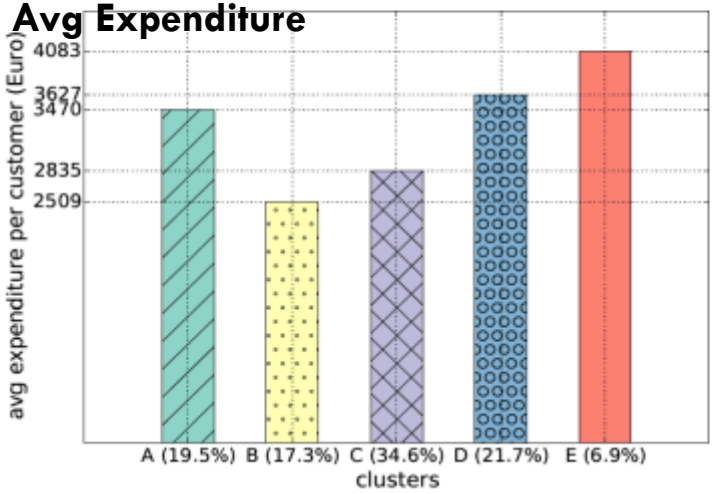
$$E(RB) = \frac{-\sum_{i=1}^{n} f(rb_i)\log f(rb_i)}{\log n}$$

Customers Classification

- Year 2012
- Leghorn province (23 shops)
- At least a shop per month
- 71,172,672 readings
- 56,448 customers
- 84,362 distinct products







Products of Systematic Customers

Vegetables and fruit are the items most shopped by the customers which are in the intersection of the systematic sets both for BRE and STRE.

Product	Sup	Product	Sup
Bananas	82.44	Fresh Eggs	64.08
Vine Tomatoes	74.22	Parsley	62.71
Sugar	72.04	Nectarines	62.55
Fennels	69.12	Green Tomatoes	62.49
Dark Zucchini	67.80	Fresh Eggs (Organic)	62.23
Bright Zucchini	67.37	Roma Tomatoes	61.49
Cherry Tomatoes	65.52	Melons	61.17

Il 60% dei bi-sistematici compra

- ZUCCHINE SCURE IT 14-21 I^ SF
- PESCHE GIALLE IT AA I^ SF
- POMODORI OBLUNGO VERDE IT 35-40 I^ SF
- POMODORI CILIEGINO IT I^ VH G 500
- POMODORI ROSSO GRAPPOLO IT I^ SF
- BANANE COOP ES 19+ I^ SF
- UOVA FRESCHE ALLEVAM A TERRA M COOP POLPA LEGNO X6
- PREZZEMOLO G. 70 CA IT MAZZI SFUSO
- PP-ZUCCHERO SEMOLATO ITALIA ZUCCHERI SCATOLA KG 1
- MELONE SEMIRETATO IT 800-1200 I^ SF
- SCONTO FIDELITY 1000 PUNTI
- RIVISTA 'NUOVO CONSUMO'

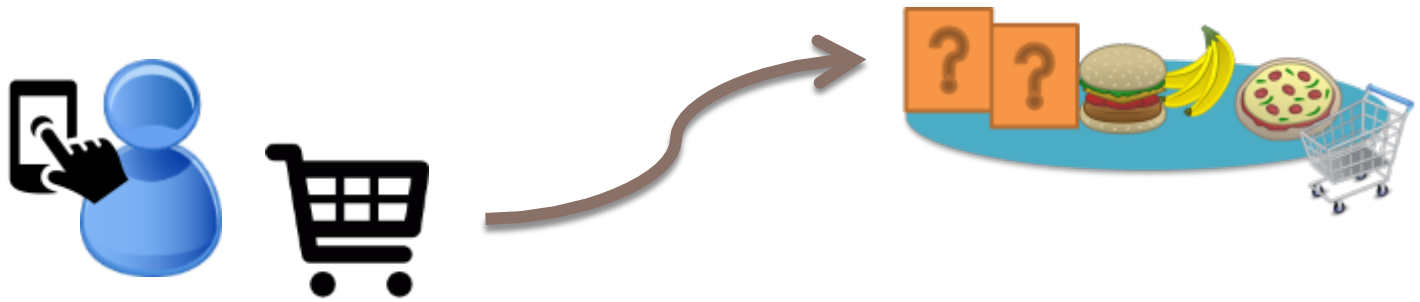
- UOVA EXTRA FRESCHE ALLEVAM A TERRA L COOP POLPA LEGNO X6
- POM.OBL.ROSSO IT30-35 I^ SF
- POM.TONDO LISCIO IT 67-82 I^SF
- ZUCCHINE CHIARE FIORE IT I^ SF
- FINOCCHI IT 8-10 I^ SF

SHOPPER COMPOSTABILE BIOFLEX ST.COOP

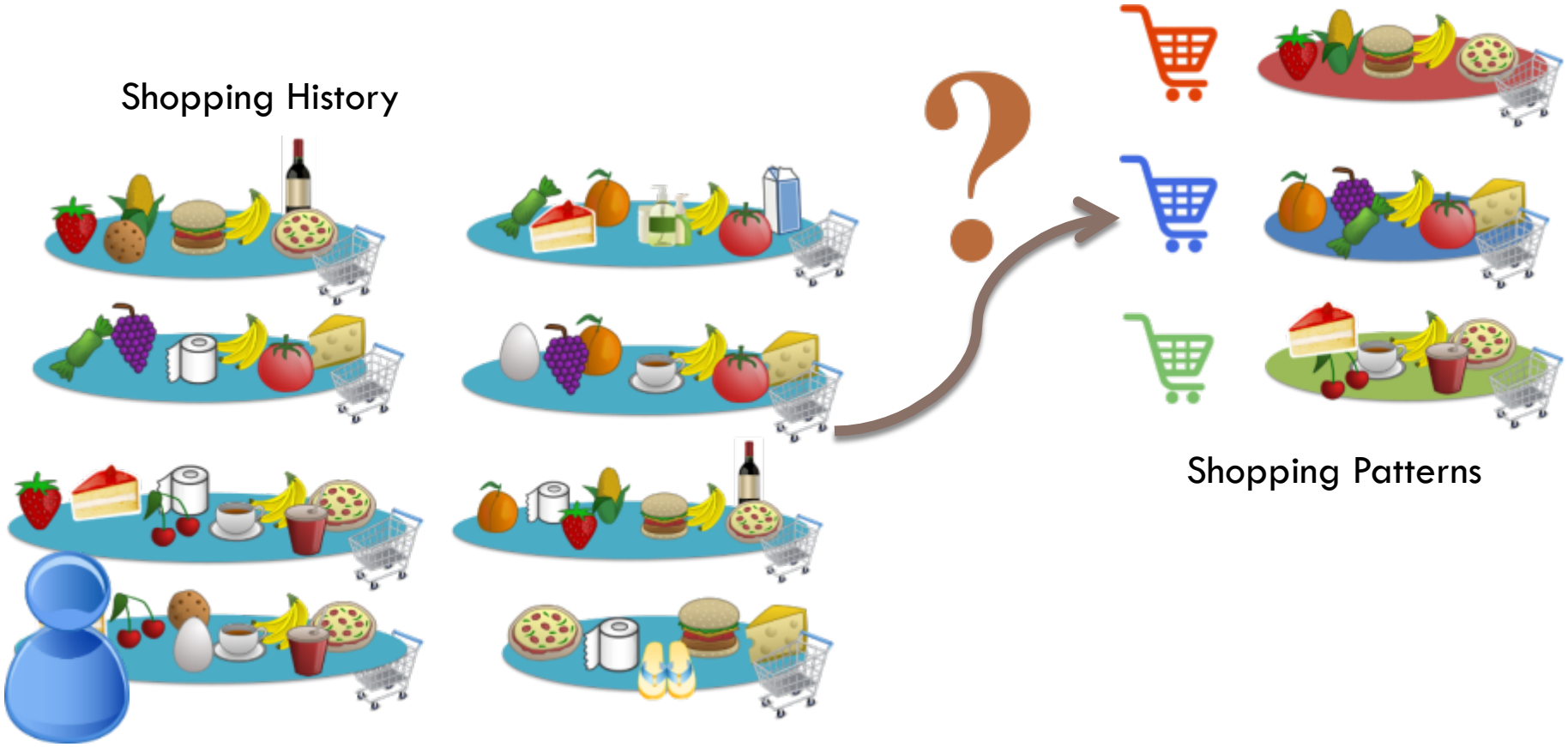
Towards a Personal Cart Assistant

Towards a Personal Cart Assistant

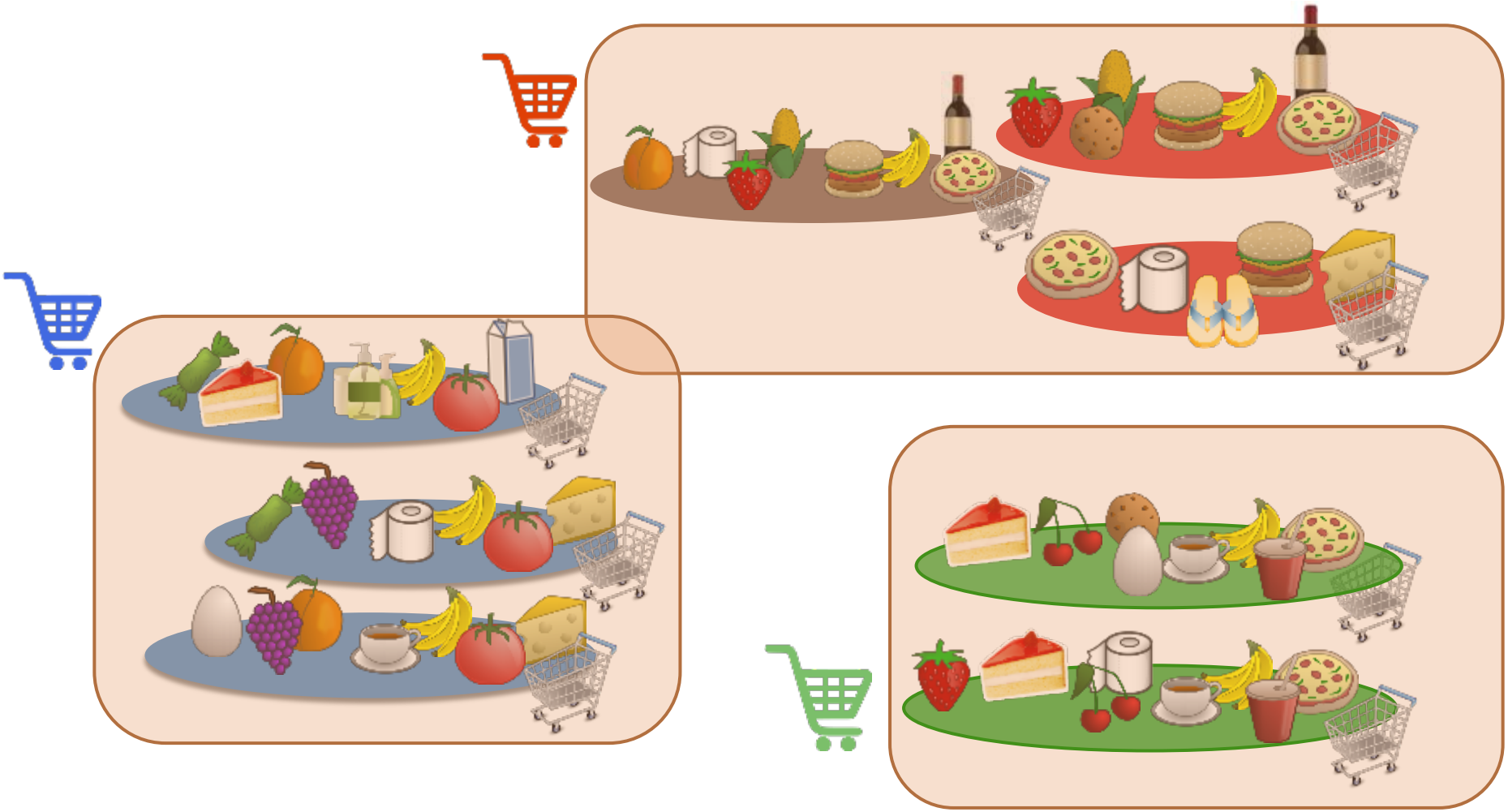
- **Goal:** Which are the products the custoemr is going to add to her current shopping list or cart?
- We need to *individually* understand which are the typical shopping patterns of each customer and exploit them to predict/suggests how the shopping list could be completed.



Typical Shopping Patterns



Autofocus Clustering Algorithm



Extract Personal Representative Baskets

Personal Cart Assistant

Preliminar Test Performances

- Improvement of 0.02 – 0.05 on the State-of-the-Art, i.e., two-three additional product correctly recommended.

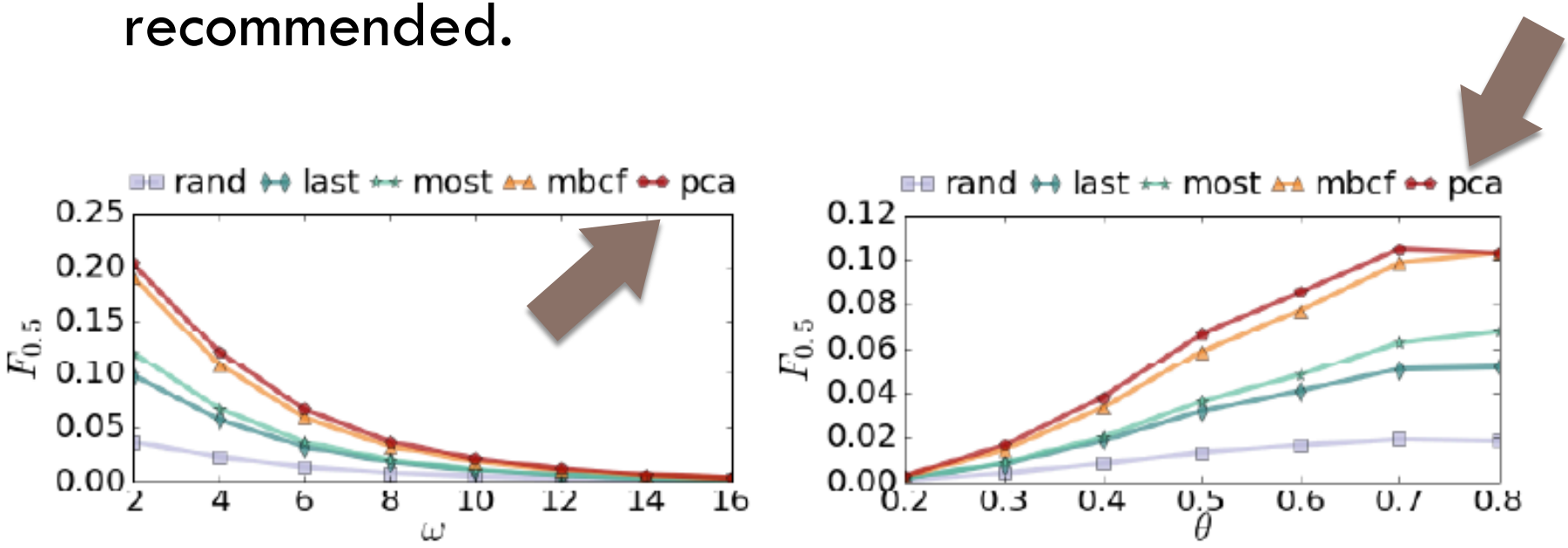


Fig. 7. Recommendation performances as $F_{0.5}$-measure on real dataset varying minimum basket length ω *(left)*, and basket split θ *(right)*.

ISTITUTO DI MANAGEMENT

Scuola Superiore Sant'Anna

Store of the future & Human dynamics from mobile data

Il gruppo di lavoro:

STUDENTI:

A. Aina
D. Bartolozzi
M. Jwad
M. Morelli
C. Viganò

AZIENDE:

Proponenti:
M. Girardelli (IBM)
C. Moiso (Telecom)
M. Caraviello (Telecom)
W. Fabbri (Unicoop)
L. Li Puma (Intesa San Paolo)
S. Menotti (SIA)

DOCENTI:

F. Giannotti (CNR)
D. Pedreschi (Univ. Pisa)

LivLab Case Study

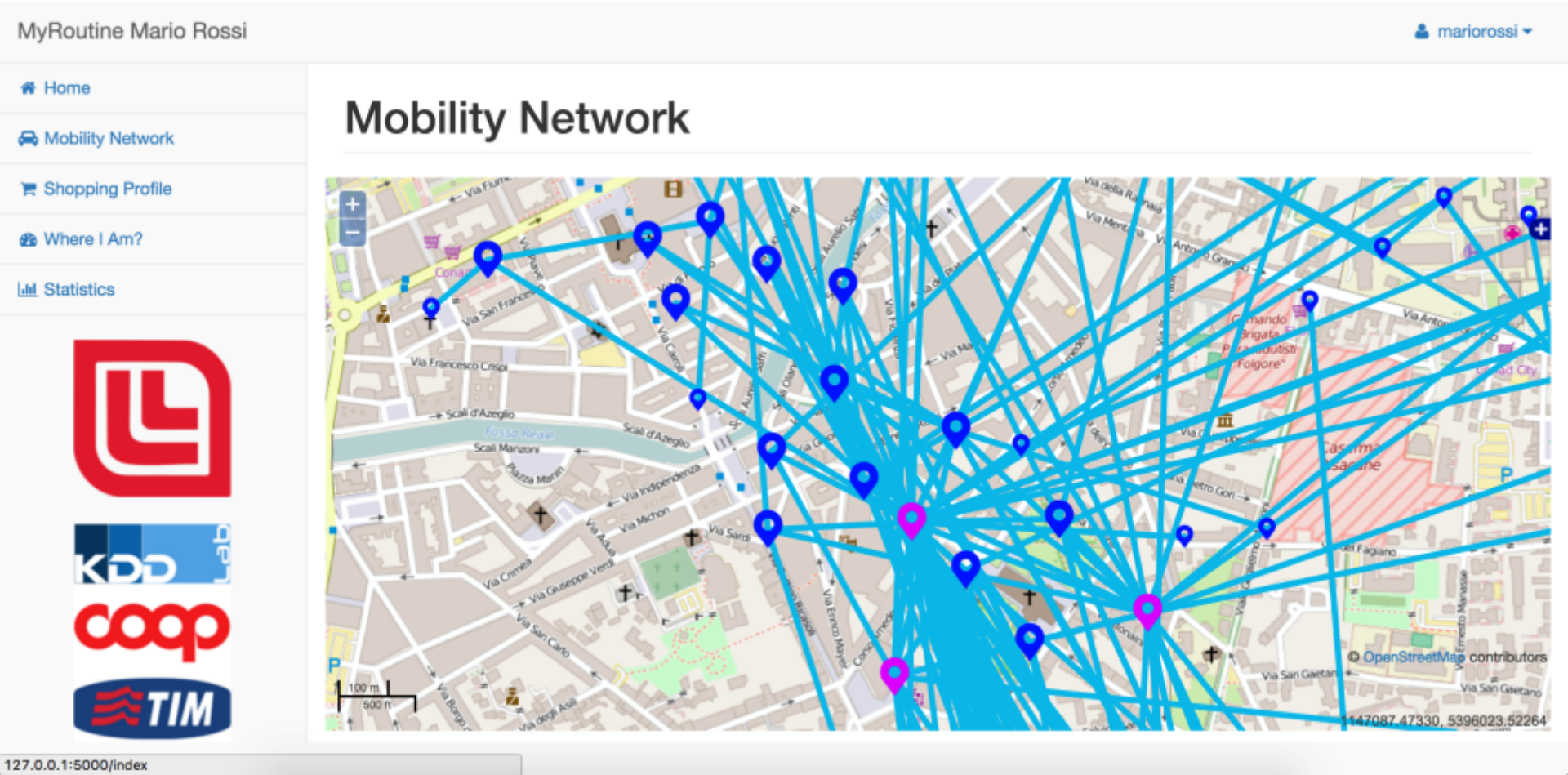
LivLab Case Study

- **Goal:** improve customers self-awareness with respect to shopping and mobility dimensions.

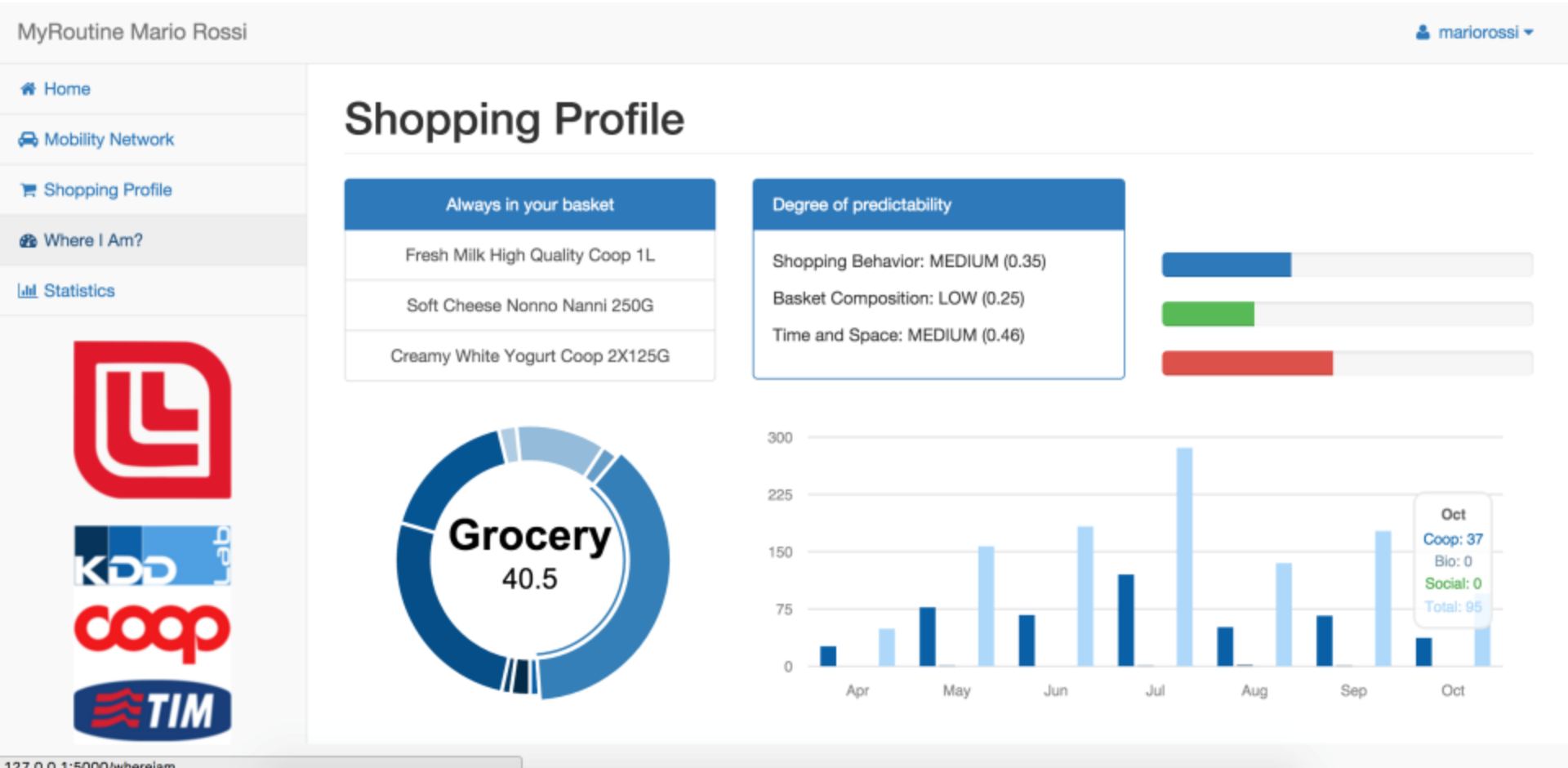
LivLab Dataset

- From 1° April 2015 to Now
- 34 Shops
- ~154 Customers (50 active???)
- ~28,600 Products
- ~9,400 Baskets
- ~400,000 Readings

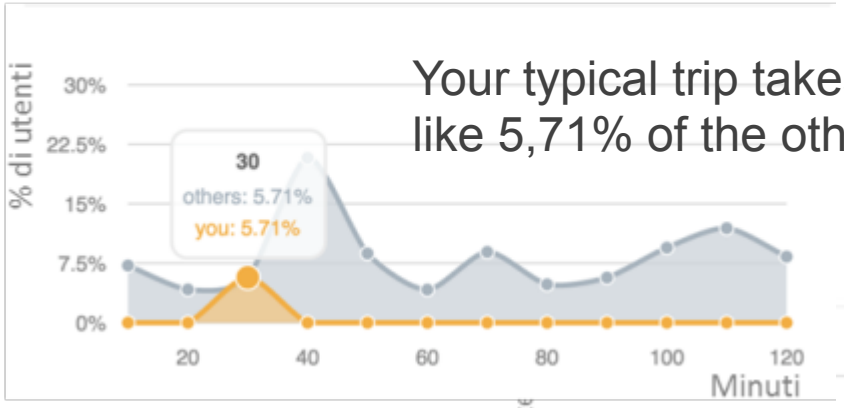
Personal Mobility Data



Personal Shopping Data

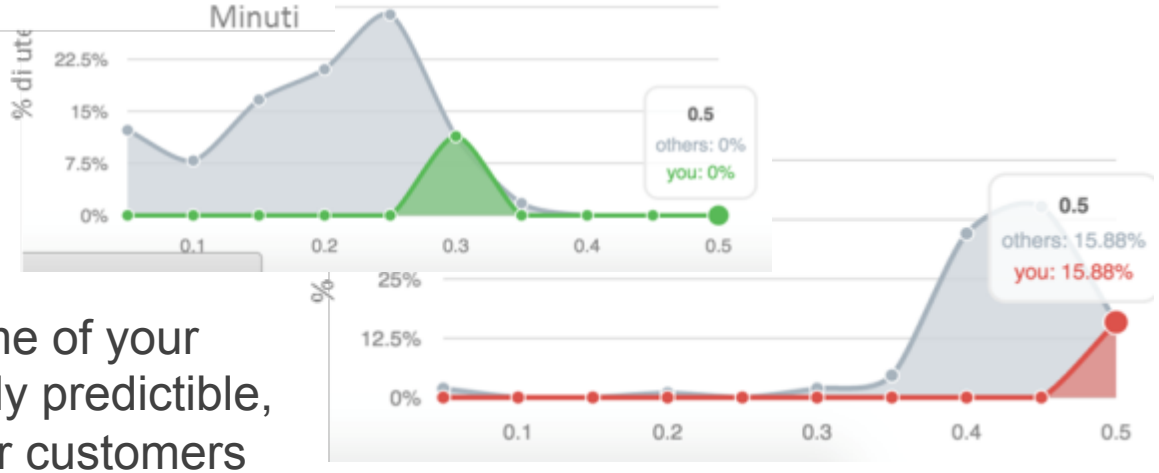


Where I Am?



Your typical trip takes 30 minutes,
like 5,71% of the other customers

Your basket compoition is quite predictible,
like 10,00% of the other customers



The shop of and the time of your
purchases are not easily predictible,
like 15,88% of the other customers

Collective Perspective

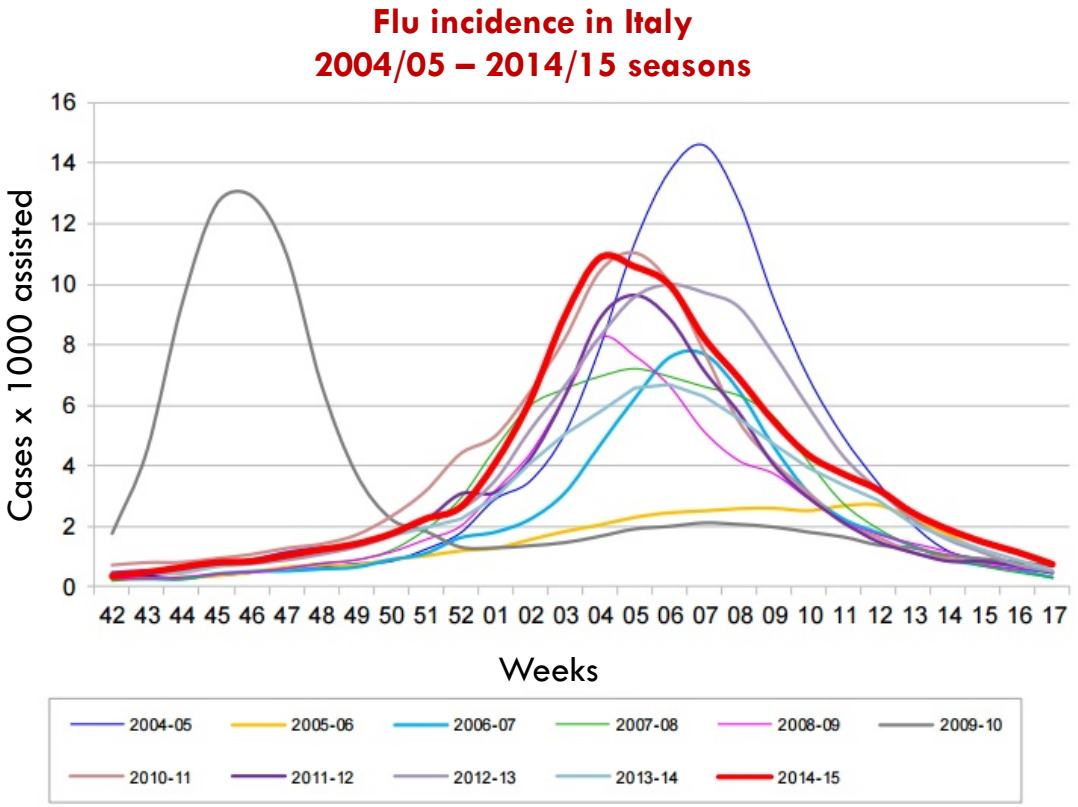
Coop Flu Trend

Coop Flu Trend

- **Goal:** Can we predict Flu by analyzing changes in purchases behaviors?
- Do different behaviors exist based on flu peaks? Do we observe changes in purchases in these periods?
- Product segments can work as proxy for prediction

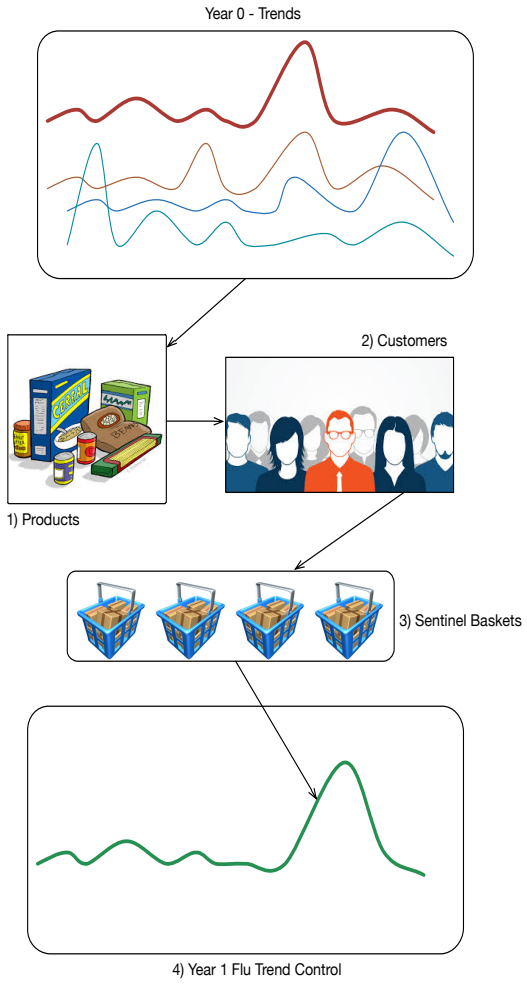
Flu Trend

- Epidemiological data from 2004/05 to 2014/15
- ~900 physicians and pediatricians
- Weekly reporting of cases of flu syndrome
- Cases divided by age group and by type of risk
- Reports from the 42nd week of the year until the last week of April of the following year



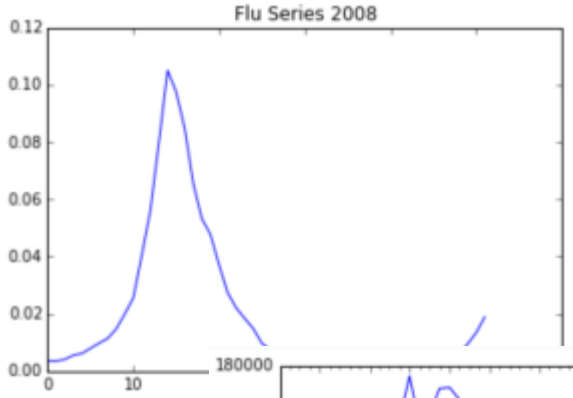
Flu Detection Workflow

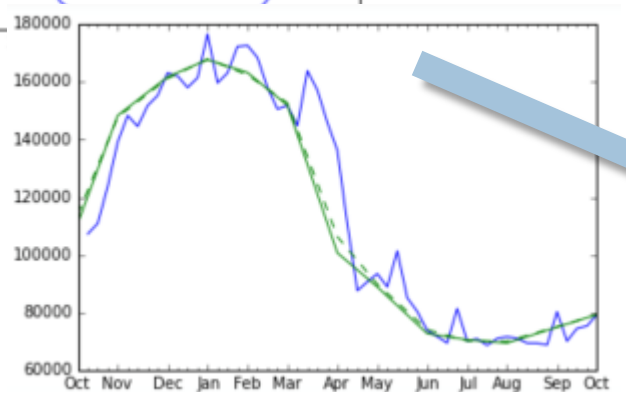
- Identify the products having adoption trend similar to the flu trend
- Identify the customers that buy them during the flu-peak
- Identify the ***sentinels***: frequent baskets of such customers during the peak
- Use the ***sentinels*** as control set for the following year flu peak

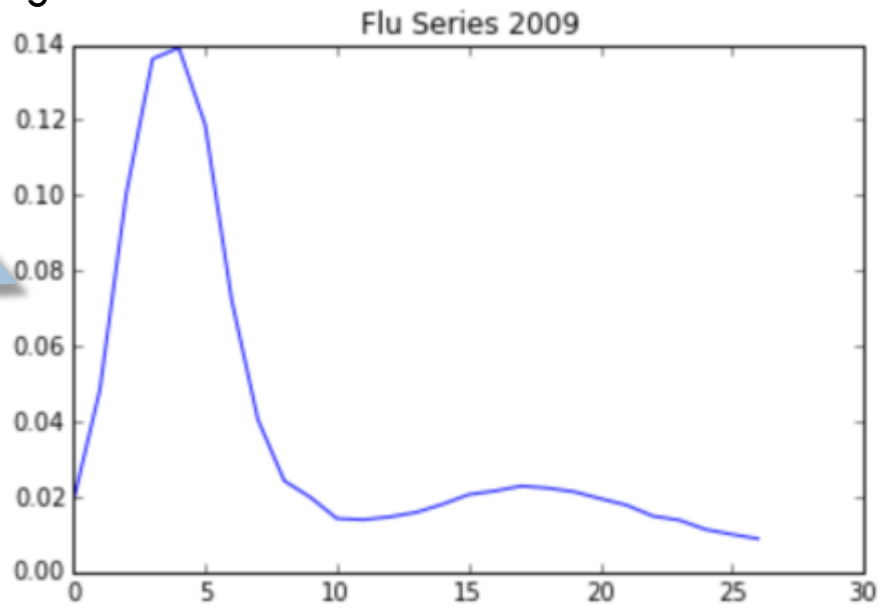


Preliminary Results

- January 2007 - August 2014
- ~50000 Active Clients
- Livorno (3 shops)
- ~3000 Products
- ~18MReadings







"IT'S A LONG WAY TO THE TOP…" PREDICTING SUCCESS VIA INNOVATORS' ADOPTIONS

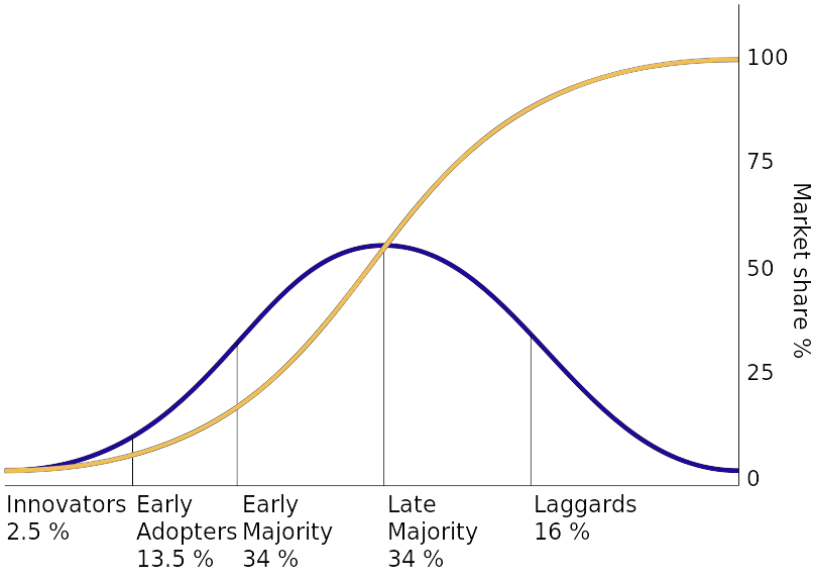
Obiettivo di analisi: esistono consumatori "speciali" che hanno una propensione ad adottare nuovi prodotti/tecnologie prima di altri? Tra questi, esiste un sottogruppo con un "sesto senso" per prodotti che avranno successo?

DOMANDA

- **E**sistono consumatori “speciali” che hanno una propensione ad adottare nuovi prodotti/tecnologie prima di altri? Tra questi, esiste un sottogruppo con un “sesto senso” per prodotti che avranno successo?

Punto di partenza

- Rogers identifica le caratteristiche distintive di ogni tipo di adopter
- Si assume che la distribuzione temporale degli Adopters segua una distribuzione normale
- Gli innovatori sono i primi 2.5% tra gli adopters
- ...ma è realmente così?



Cosa significa “raggiungere” il successo?

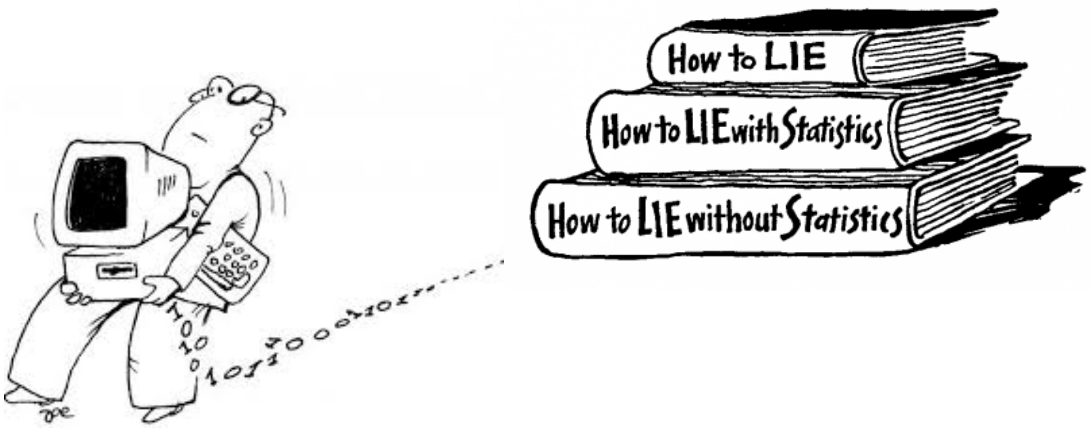
A **Data Driven** definition

Data tell no lies!

(…hopefully…)

Idea:

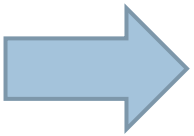
- Identificare diversi pattern temporali di adozione
- Correliamoli al volume dell’adozione



Data Preparation

- Per studiare le adozioni, occorre modellare i dati con serie temporali

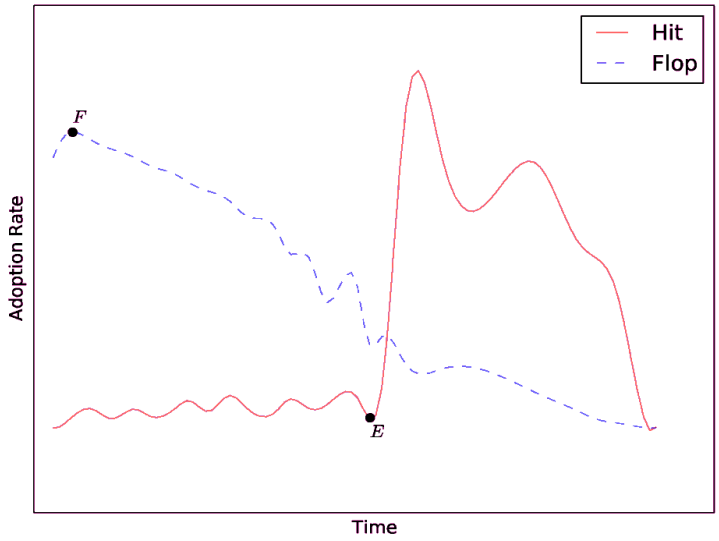
Prodotto	Cliente	Unità temporale
1	A	U1
1	B	U2
1	C	U1
2	C	U2
2	D	U2
2	A	U4
2	E	U4



Prodotto	Serie Temporale
1	<U1,2> , <U2,1>
2	<U1,1>,<U2,2>,<U4, 2>

Hit and Flop: definizioni

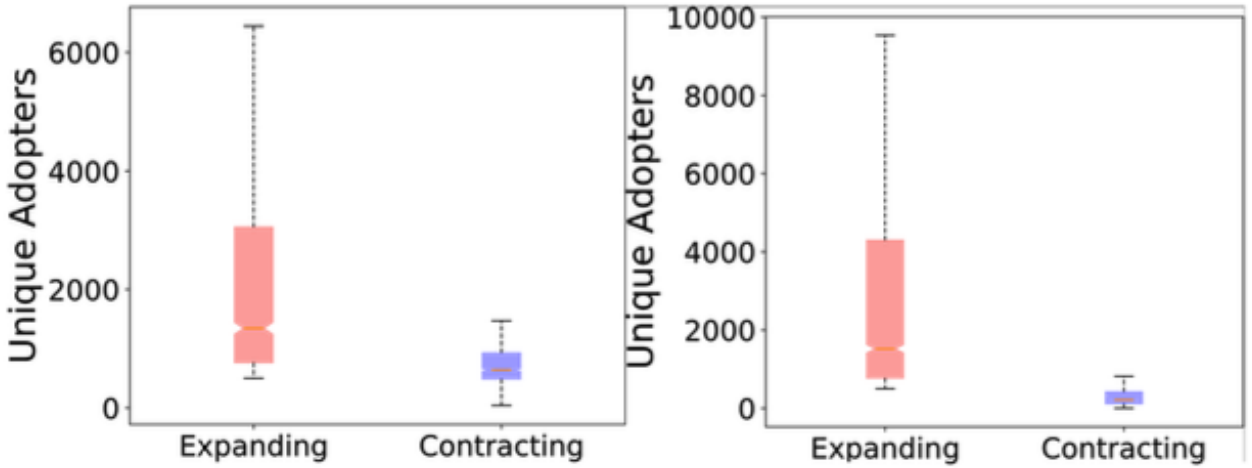
- Un **Hit** è un bene il cui trend di adozione cresce lentamente nel tempo fino a raggiungere un punto di esplosione (E nella figura), che segna l'inizio di un forte aumento del numero di adozioni.
- Un **Flop** è un bene il cui trend di adozione non cresce in maniera considerevole nel tempo, oppure raggiunge subito un punto di massimo (F in figura) per poi decrescere rapidamente.



- Non si considerano i volumi di vendita, ma i trend relativi
- La definizione di "successo" non corrisponde a quello in senso comune (un nuovo iPhone cadrebbe nella nostra definizione di Flop)

…confermati dal Volume

Trends classes and their volume of adoption



Successful trends (Hits).
The ones describing an increase of adoption rate through time capturing an expansion of individual items' adopter base

(In our analysis: 30% Artists in Last.fm, 40% products in Coop)

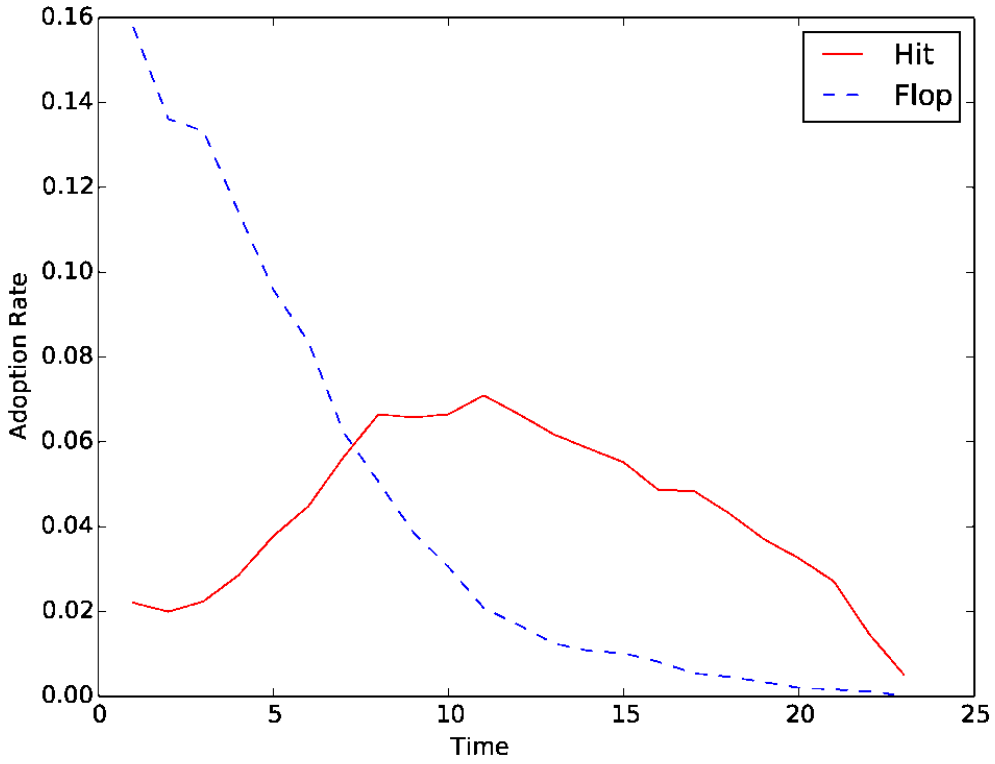
Unsuccessful trends (Flops).
The ones in which the adoption rate do not increase considerably over time or even reach an early maximum only to start to decrease rapidly

Il Metodo

- Si estraggono i profili dei beni Hit e Flop
- Si individuano gli innovatori
- Si calcolano gli indicatori di successo/insuccesso per tutti gli innovatori
- Si raffinano e si consolidano gli insiemi di indicatori positivi/negativi
- Si definisce una tecnica di predizione rule-based utilizzando gli indicatori calcolati in precedenza

Sperimentazione

- Una volta costruite le serie temporali per tutti i prodotti nuovi, queste vengono clusterizzate



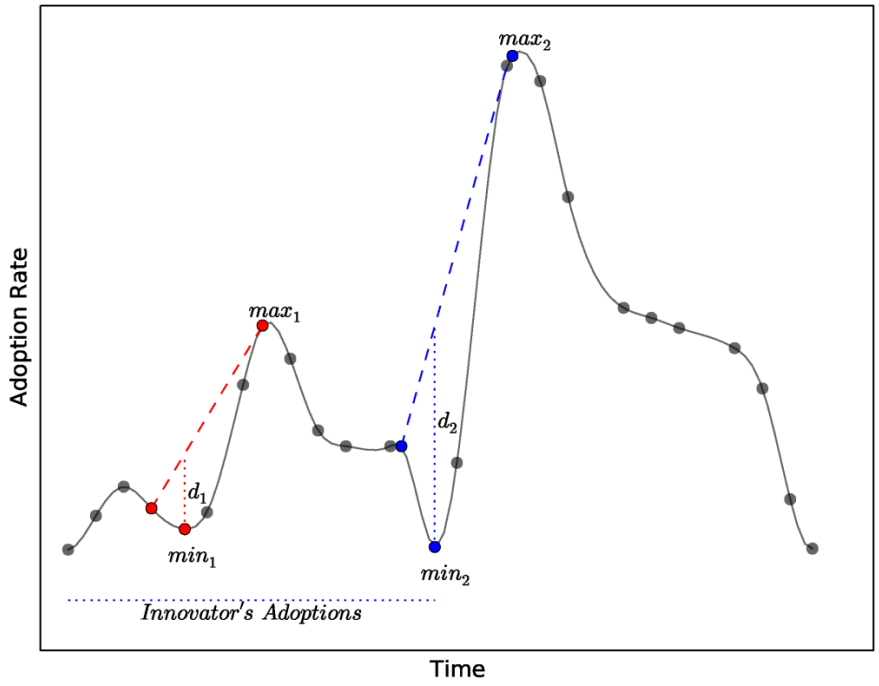
Clustering: K-means

K=2 con SSE

Unità temporale: settimana

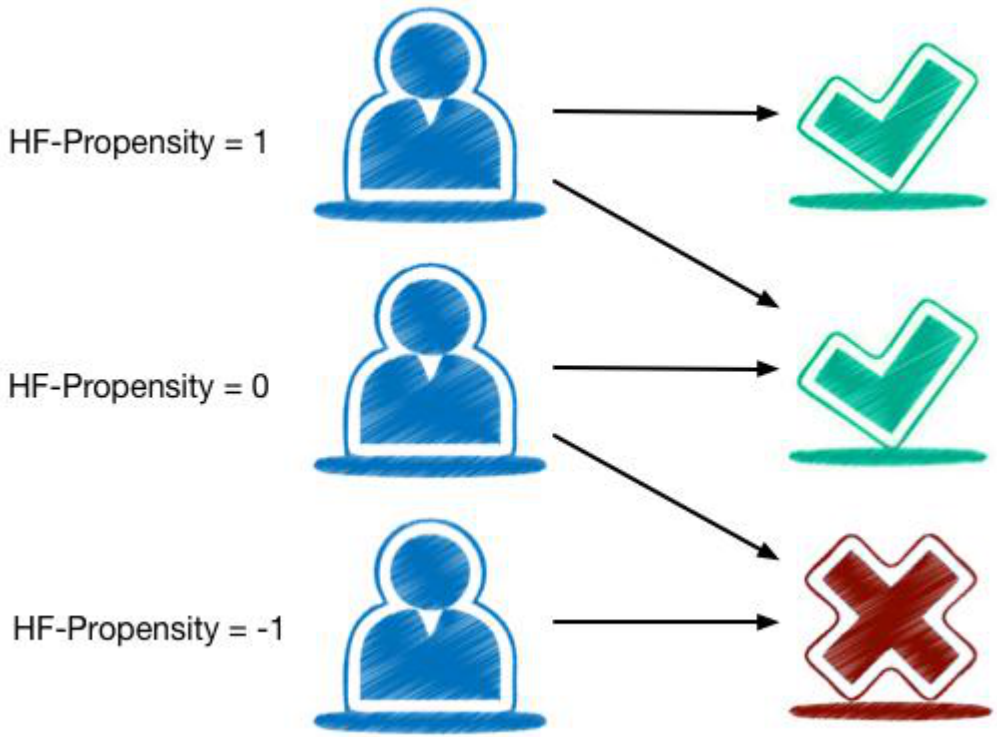
Identificazione Innovatori

- Si identificano i minimi locali (min_1 e min_2) e i massimi locali (max_1 e max_2)
- Si calcolano le distanze d_1 e d_2
- Si sceglie min_2 come punto discriminante, in quanto $d_2 > d_1$
- Si definiscono innovatori tutti i clienti che hanno adottato il bene prima di min_2



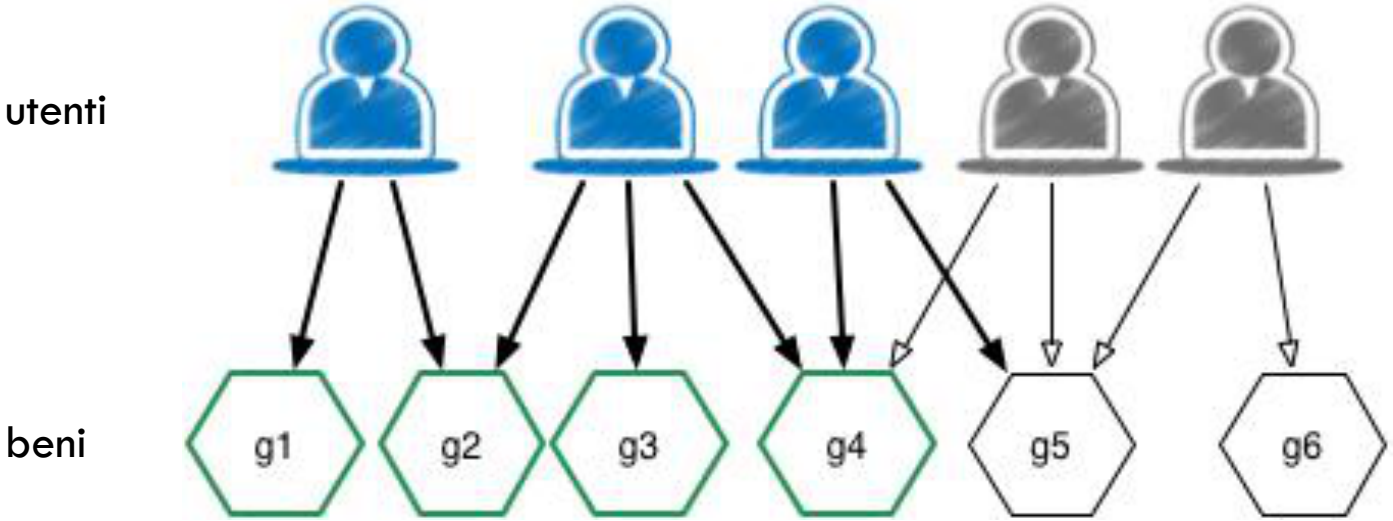
Propensione degli innovatori per i beni Hit e Flop

Si misura, per ogni cliente, la sua propensione ad adottare beni Hit o Flop, assegnando un +1 se ha adottato un Hit da innovatore, un -1 se ha adottato un Flop, nessun incremento/ decremento se ha adottato un Hit dopo l'esplosione



Identificazione degli Hitters e dei Floppers

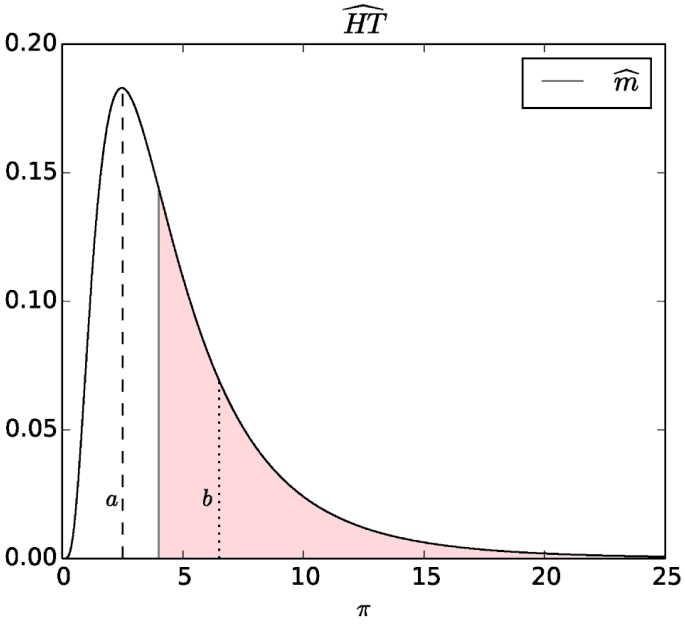
- Bisogna ridurre l'insieme di clienti in modo da filtrare solo gli indicatori forti per gli Hit ed i Flop
- Si costruiscono due grafi bi-partiti <cliente,prodotto> (uno per beni Hit ed uno per beni Flop), dove un arco tra un nodo cliente ed un nodo prodotto viene tracciato se quel cliente ha adottato quel prodotto
- Si risolvono i due problemi di Copertura Minima Pesata (con HF-propensity) dei prodotti
- I clienti selezionati coprono collettivamente almeno una frazione α dell'insieme dei beni ed almeno una frazione β degli archi entranti (supporto) dei beni

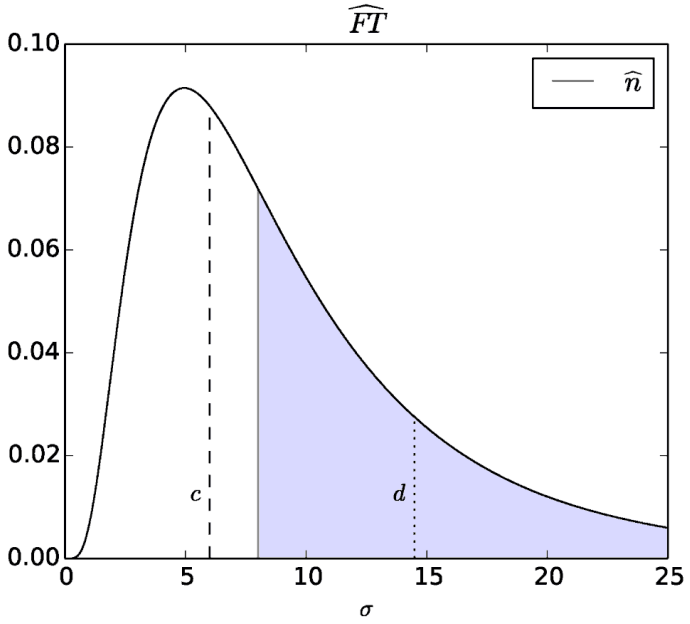


Esempio con
$\alpha = \beta = 0.6$

Modello predittivo

- Si tracciano le distribuzioni del numero di Hitter (sx) e Flopper (dx) che hanno adottato i prodotti
- Si tracciano le mediane delle funzioni
- Per ogni prodotto g, si calcolano $\pi(g)$ (#Hitters che hanno adottato g) e $\sigma(g)$ (#Floppers che hanno adottato g)





$\pi(g)$ =a $\sigma(g)$=d **Flop**

$\pi(g)$ =b $\sigma(g)$=c **Hit**

$\pi(g)$=b $\sigma(g)$=d **Flop** (calcolo distanze)

$\pi(g)$=a $\sigma(g)$=c **?**

Analisi sperimentale

- 5.605 articoli
- 620.026 clienti
- 11.204.984 clienti
- Periodo di un anno

- 3.749 Hits
- 1.856 Flops

Risultati

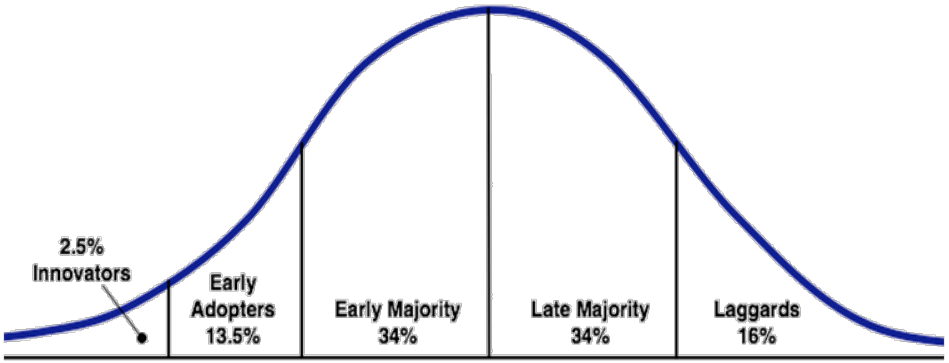
- Il modello viene confrontato con
  - Null Model (tutti i clienti comprano in un istante random)
  - ER-H&F: la fase di identificazione degli innovatori è sostituita considerando la soglia di Rogers (2.5%)
  - ER: la predizione viene fatta utilizzando soltanto gli innovatori estratti secondo la soglia del 2.5%

Coop	H&F		ER-H&F		ER		Null Model	
	avg	*std*	*avg*	*std*	*avg*	*std*	*avg*	*std*
Precision	0.781	0.91	0.825	0.211	0.00	0.00	0.547	0.010
NPV	0.316	0.121	0.384	0.057	0.292	0.00	0.051	0.028
Recall	0.586	0.292	0.031	0.017	0.00	0.00	0.818	0.043
Specificity	0.522	0.367	0.983	0.019	1.00	0.00	0.361	0.024

Adopters: Innovators

- **Diffusion of Innovations**
  [Rogers 1962]

- Five "category" of **Adopters** based on the time of first adoptions:
  - Each one has its own semantics;
  - Temporal distribution
    Assumed to be a Gaussian;
  - Categories proportion is univocally determined
    (i.e. Innovators are <u>always</u> the first 2.5%)



Goods: Hits and Flops

- Retail market products,
- Music Artists,
- Business and stores…

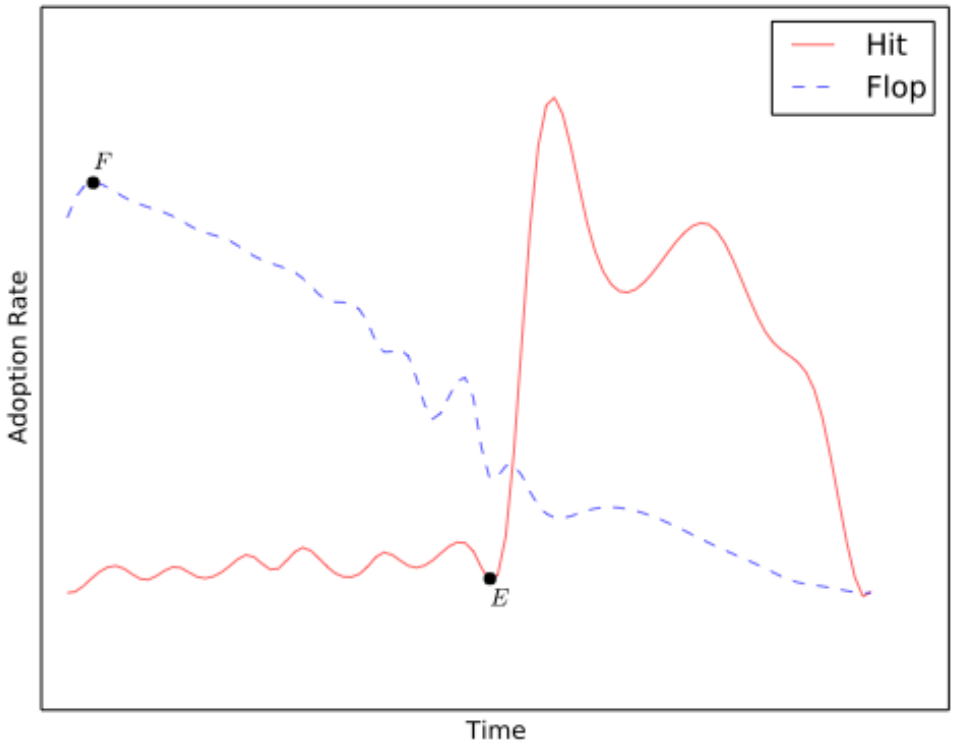
What is a successful (Hit) good?

And an unsuccessful (Flop) one?

Hits and Flops share the same set of adopters or not?

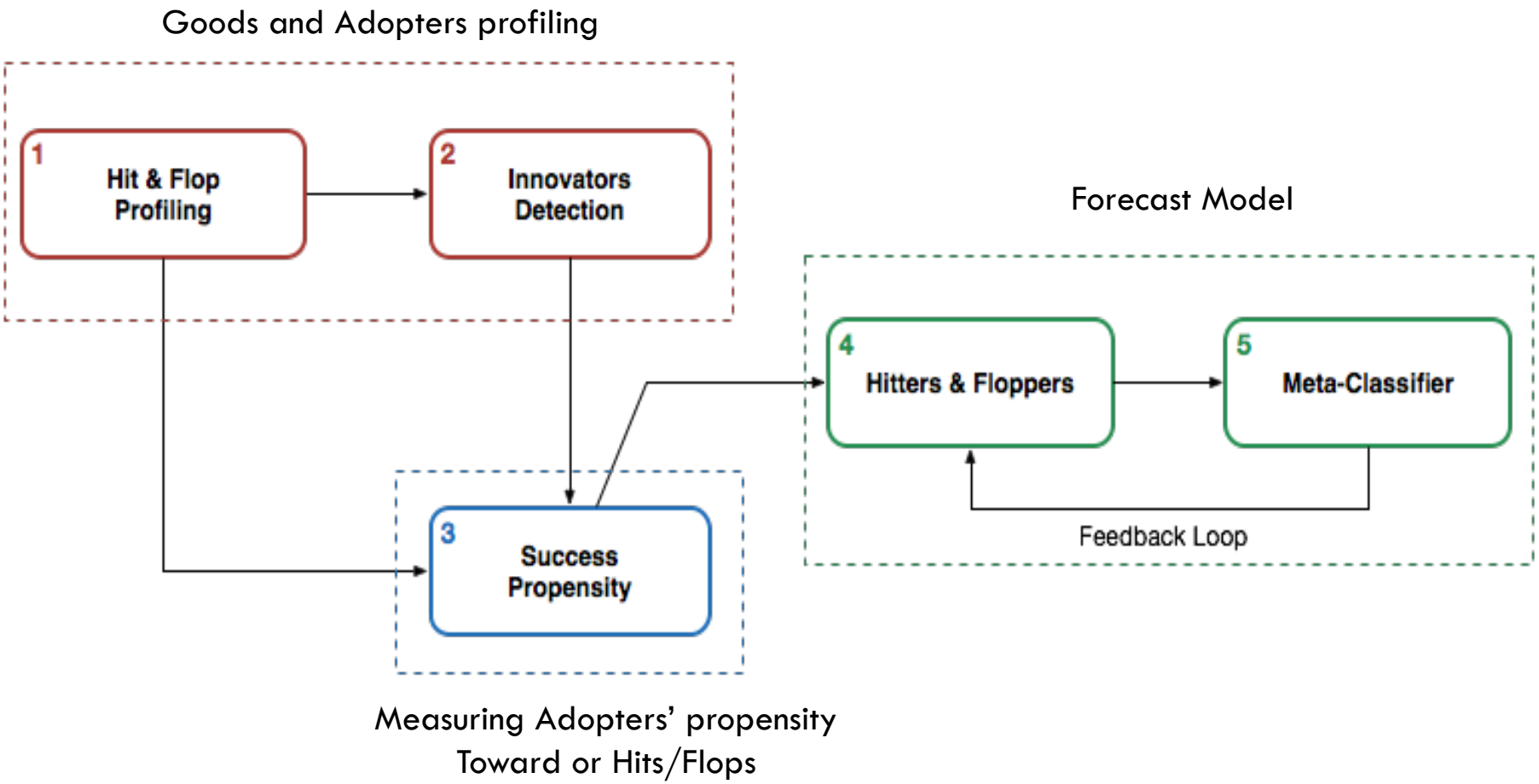
Hits & Flops: qualitative definitions

- **Hit**
  - A good whose trend slowly increases trough time until reaching an explosion point that marks the start of a sharp rising of its adoptions.
- **Flop**
  - A good whose adoption trend does not increase considerably over time or even reaches an early maximum only to sharply decrease.



Given a **partial observation** of the **adoptions** of a **novel good** can we decide if it will became an **Hit** or a **Flop**?

Hit&Flop: Workflow



Forecast Evaluation*

COOP	H&F	ER-H&F	ER	NM
PPV	**.781**(.09)	.825(.21)	0(0)	.547(.01)
NPV	.316(.12)	.384(.06)	.292(0)	.05(.03)
Recall	**.586**(.29)	.03(.01)	0(0)	.818(.04)
Specificity	.522(.38)	.982(.02)	1(0)	.361(.02)

Last.fm	H&F	ER-H&F	ER	NM
PPV	**.766**(.03)	.290(.37)	0(0)	.644(0)
NPV	**.471**(.04)	.047(.39)	.351(0)	.026(.04)
Recall	.520(.04)	.006(.01)	0(0)	.990(.02)
Specificity	.727(.06)	.970(.02)	1(0)	.007(.01)

Yelp	H&F	ER-H&F	ER	NM
PPV	**.990**(.01)	1(0)	0(0)	.488(.04)
NPV	**.631**(.17)	.341(.11)	.306(0)	.099(.08)
Recall	**.897**(.09)	.654(.11)	0(0)	.933(.01)
Specificity	**.906**(.10)	1(0)	1(0)	.007(.01)

Datasets

Dataset	Goods	Adopters	Adoptions	Period	Obs. window
COOP	5605	620026	11204984	1 year	4 weeks
Last.fm	1806	50837	882845	2 years	2 months
Yelp	2499	141936	427894	10 years	30 months

Competitors

H&F: Hits&Flops

ER-H&F: Hits&Flops with Roger's Innovators

ER: Rogers's Innovators

NM: Hits&Flops on Null Model (avg. 100 models)

Results in a nutshell

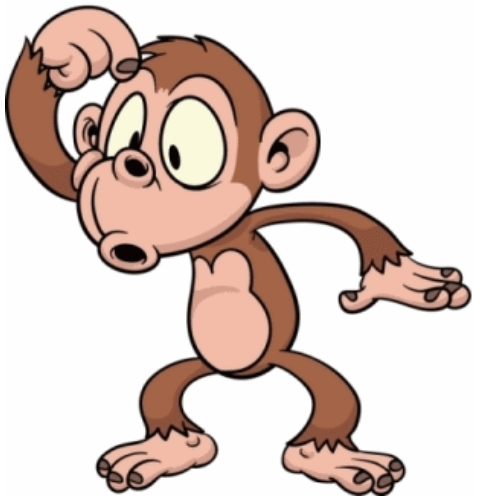
- H&F guarantee the most stable predictive performances in terms of PPV and Recall
- ER is not able to provide useful classification (2.5% fixed innovator threshold)
- ER–H&F suffer the constrains imposed by ER

*Results after a 10-fold cross validation

Statistical Significance Analysis

(Null model evaluation)

We have assumed that adopters **do not** act like monkeys… but what would happen if they **did** so?



Null Model:

- Maintain adoption trends;
- Maintain temporal volume of adoptions per adopter;
- Destroy the adopter's choice of items (i.e. adopters act like chimps)

Statistical Significance:

Are the results of H&F driven by its ability of extracting information from real data?
To what extent randomness can explain them?

z-test

z-values*

	PPV	NPV	Recall	Specificity
COOP	23,4	9,5	5,4	10,9
Last.fm	61	9,7	24,2	48
Yelp	14,3	6,8	12	69,2

* z-value > 3.9 ⇒ p-value = 0

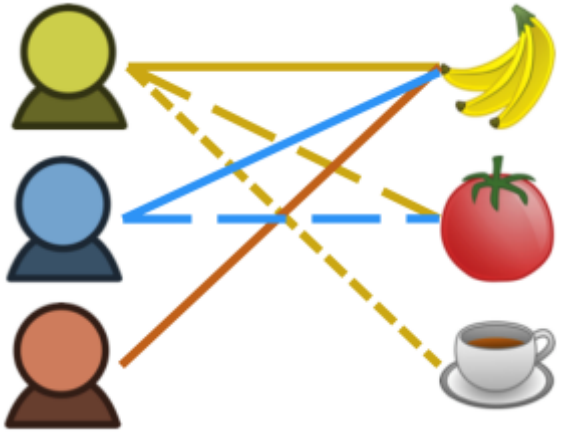
Nowcasting GDP & Well-Being

Nowcasting GDP & Well-Being

- **Goal:** Estimating well-being observing customers retails through a measure called sophistication.
- GDP (Gross Domestic Product): is the market value of all goods and services produced within a country in a given period of time.
- GDP is thought to capture average prosperity
- Can we estimate GDP? Can we nowcast it?

Customers–Products & Sophistication

- ***G* = *(C, P, E)*** is a bipartite network
- **C** is the set of customers
- ***P*** is the set of products
- ***E*** is the set of edges **e** = ***(i, j, w_ij)*** ***w_ij*** is the number of times customer ***i*** bought product **j**
- Customers are sophisticated if they purchase sophisticated products.
- Products are sophisticated if they are bought by sophisticated customers.



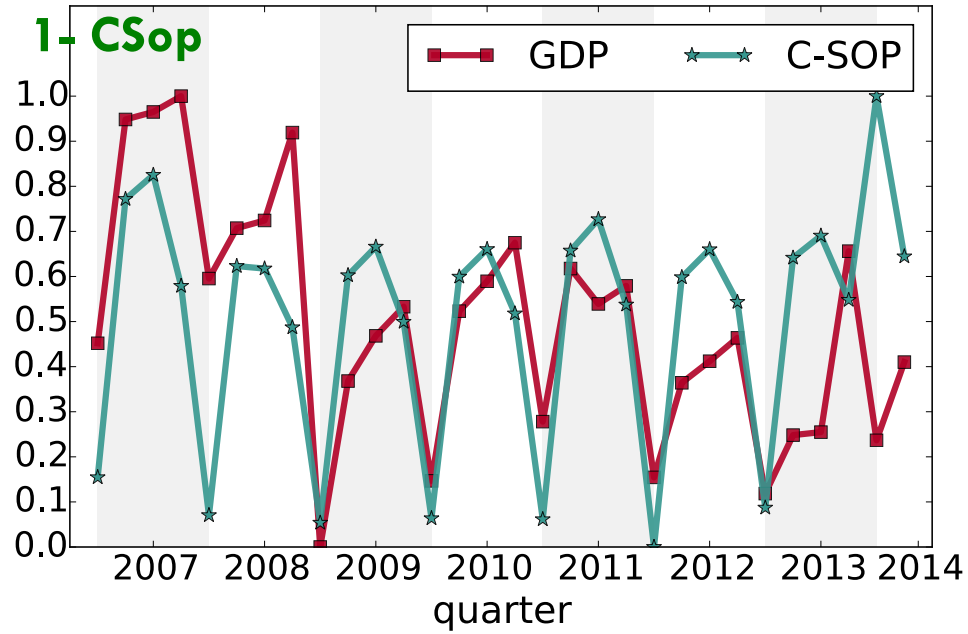
$$c_i^{(n)} = \sum_{j=1}^{|P|} \frac{1}{k_j} M_{ij} p_j^{(n-1)}$$

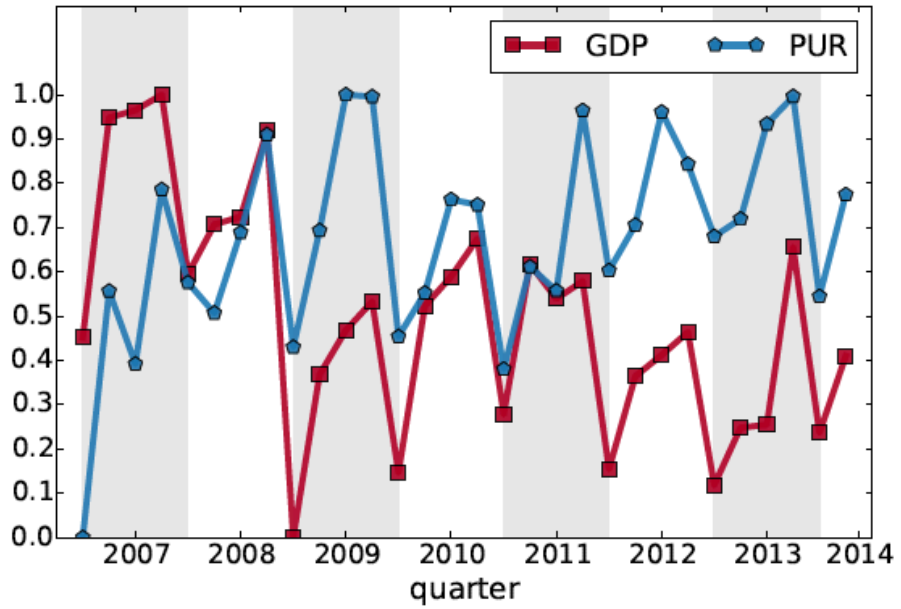
$$p_j^{(n)} = \sum_{i=1}^{|D|} \frac{1}{k_i} M_{ij} c_i^{(n-1)}$$

GDP vs C-SOP – GDP vs P-SOP

- We considered the whole dataset w.r.t. the segment marketing hierarchy
- Relation between GDP and customers sophistication (left) and product purchased (right)
- Correlation passes for p > 0.01

-2 shift





Products Sophistication

SOP Rank	Product
1	Cosmetics
2	Underwear for man
3	Furniture
4	Multimedia service
5	Toys
…	…

SOP Rank	Product
…	…
-5	Fresh Cheese
-4	Red Meat
-3	Spaghetti
-2	Bananas
-1	Short Pasta

Products Sophistication

SOP Rank	Product
1	Cosmetics
2	Underwear for man
3	Furniture
4	Multimedia service
5	Toys
…	

SOP Rank	Product
…	…
-5	Fresh Cheese
-4	Red Meat
-3	Spaghetti
-2	Bananas
-1	Short Pasta