Explaining the Product Range Effect in Purchase Data

Diego Pennacchioli, Michele Coscia, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti

diego.pennacchioli@isti.cnr.it

Institutions Markets Technologies IMT INSTITUTE FOR ADVANCED STUDIES LUCCA

KDD Lab

Center for International Development at Harvard University

IN SUPREMÆ DIGNITATIS 1343

2013 IEEE International Conference on Big Data, Santa Clara - CA, 6-9/10/2013

Outline

- the purchase behavior - issues in data analysis
- new measures to represent the personal utility function
- introducing the range effect
- conclusions

What drives customer behavior?

What drives customer behavior?

generic utility function (rationality)

What drives customer behavior?

generic utility function (rationality)

personal utility function (diversity)

What about personal utility function?

What about personal utility function?

- how often and in which quantity do I need that good?

What about personal utility function?

- how often and in which quantity do I need that good?

- how much do I desire that good?
- how useful is that good for me?
- how "unique" is that good? (can I replace it with other goods satisfying the same need?)

What about personal utility function?

QUANTITY PURCHASED

SOPHISTICATION DEGREE

Santa Clara, October 7th 2013 - Diego Pennacchioli, National Research Council of Italy (Pisa) - diego.pennacchioli@isti.cnr.it

What is sophistication?

What is sophistication?

- neither the price

What is sophistication?

- neither the price

- nor the functional feature of the product

What is sophistication?

- the sophistication of a good is strongly related to the need that the good satisfies

What is sophistication?

- the sophistication of a good is strongly related to the need that the good satisfies

What is sophistication?

- the sophistication of a good is strongly related to the need that the good satisfies

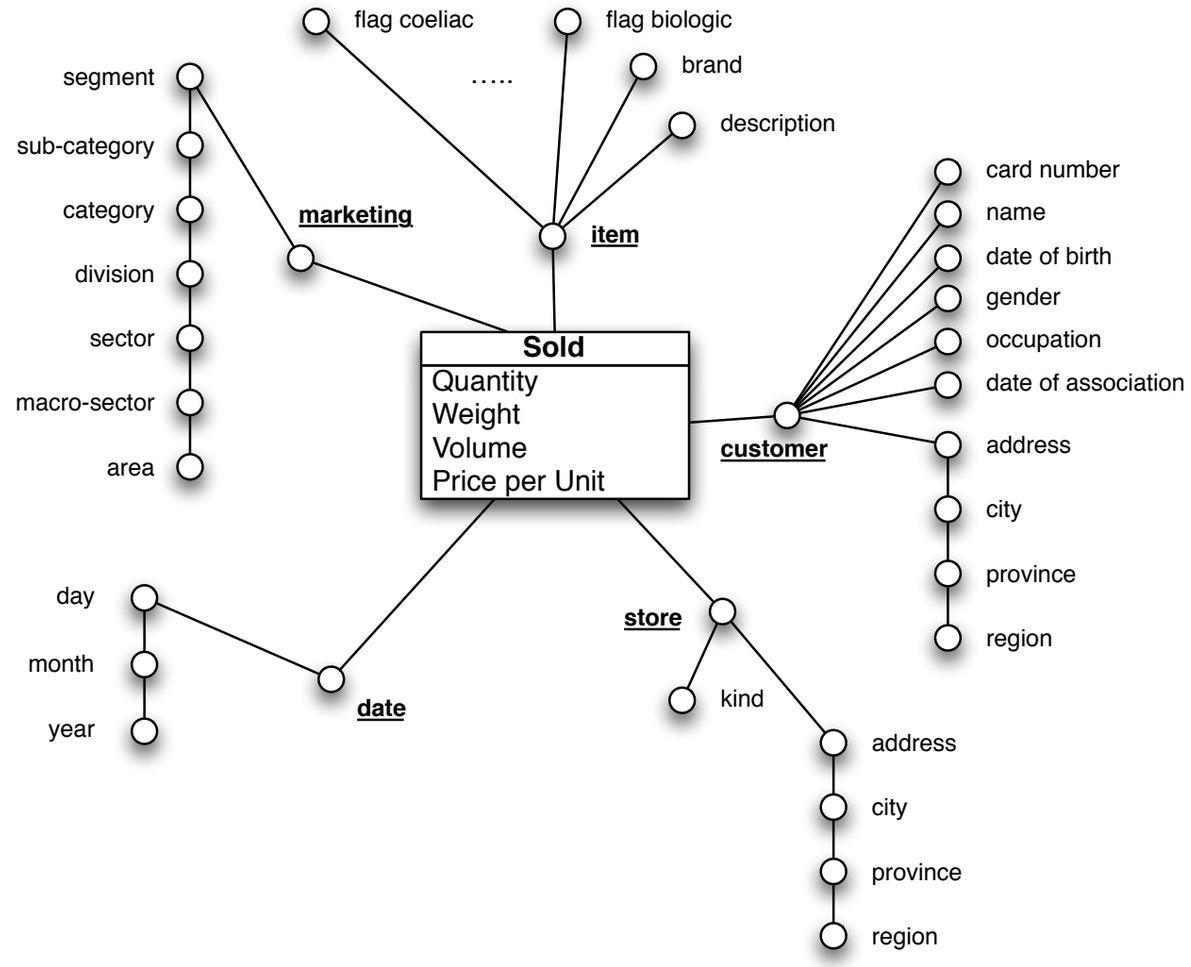
- to be considered "sophisticated" a product needs to satisfy two constraints:
  - it has to be sold to few customers
  - the customers buying it have to buy all products that are less sophisticated than it

Data

- largest italian retail distribution Company
- time window: 2007-01-01 / 2012-12-31
- 1,066,020 active and recognizable customers
- 138 stores over the whole Italy's west coast
- 345,208 different items
- 2 Billions purchases

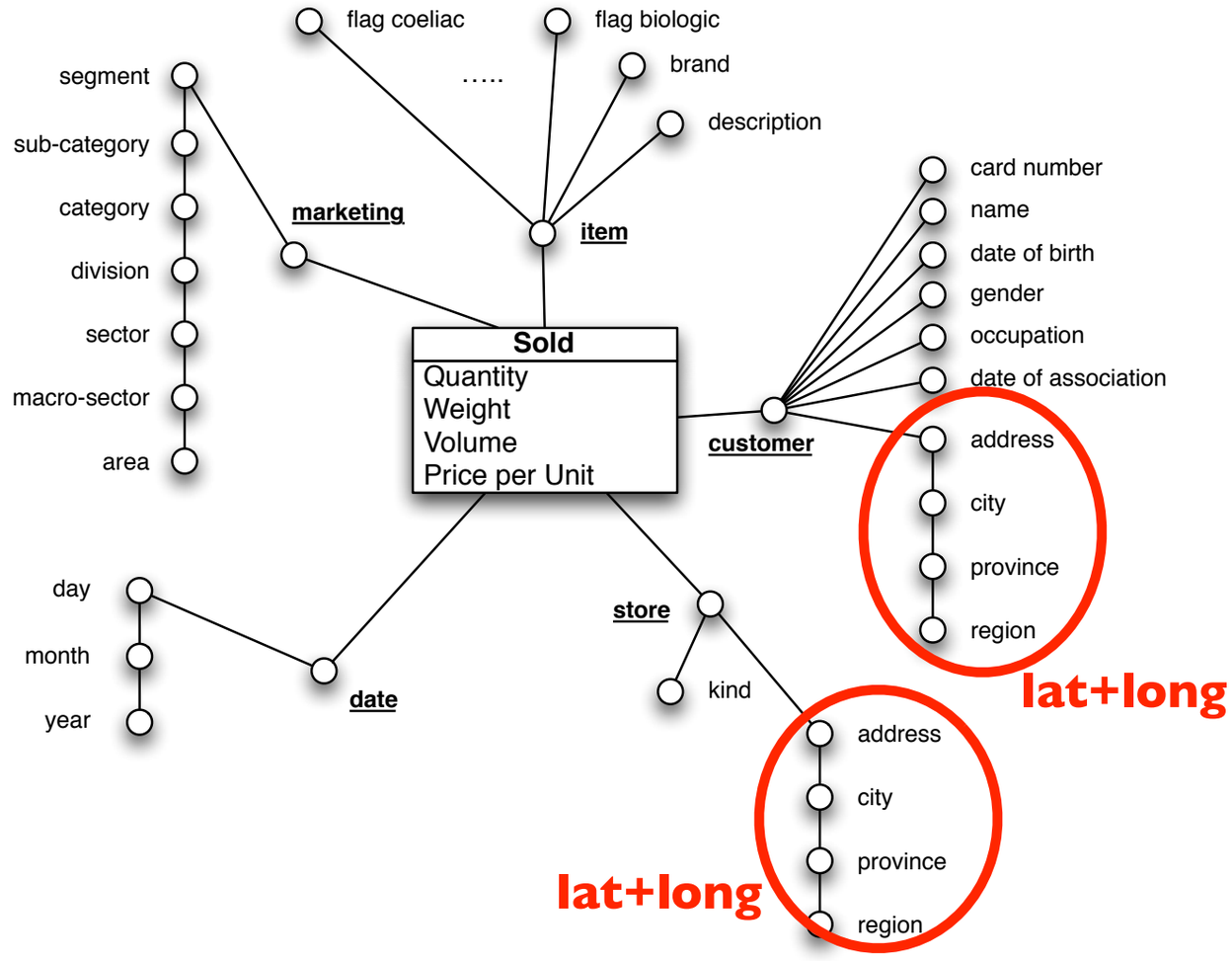
Data

- largest italian retail distribution Company
- time window: 2007-01-01 / 2012-12-31
- 1,066,020 active and recognizable customers
- 138 stores over the whole Italy's west coast
- 345,208 different items
- 2 Billions purchases



Data

- largest italian retail distribution Company
- time window: 2007-01-01 / 2012-12-31
- 1,066,020 active and recognizable customers
- 138 stores over the whole Italy's west coast
- 345,208 different items
- 2 Billions purchases

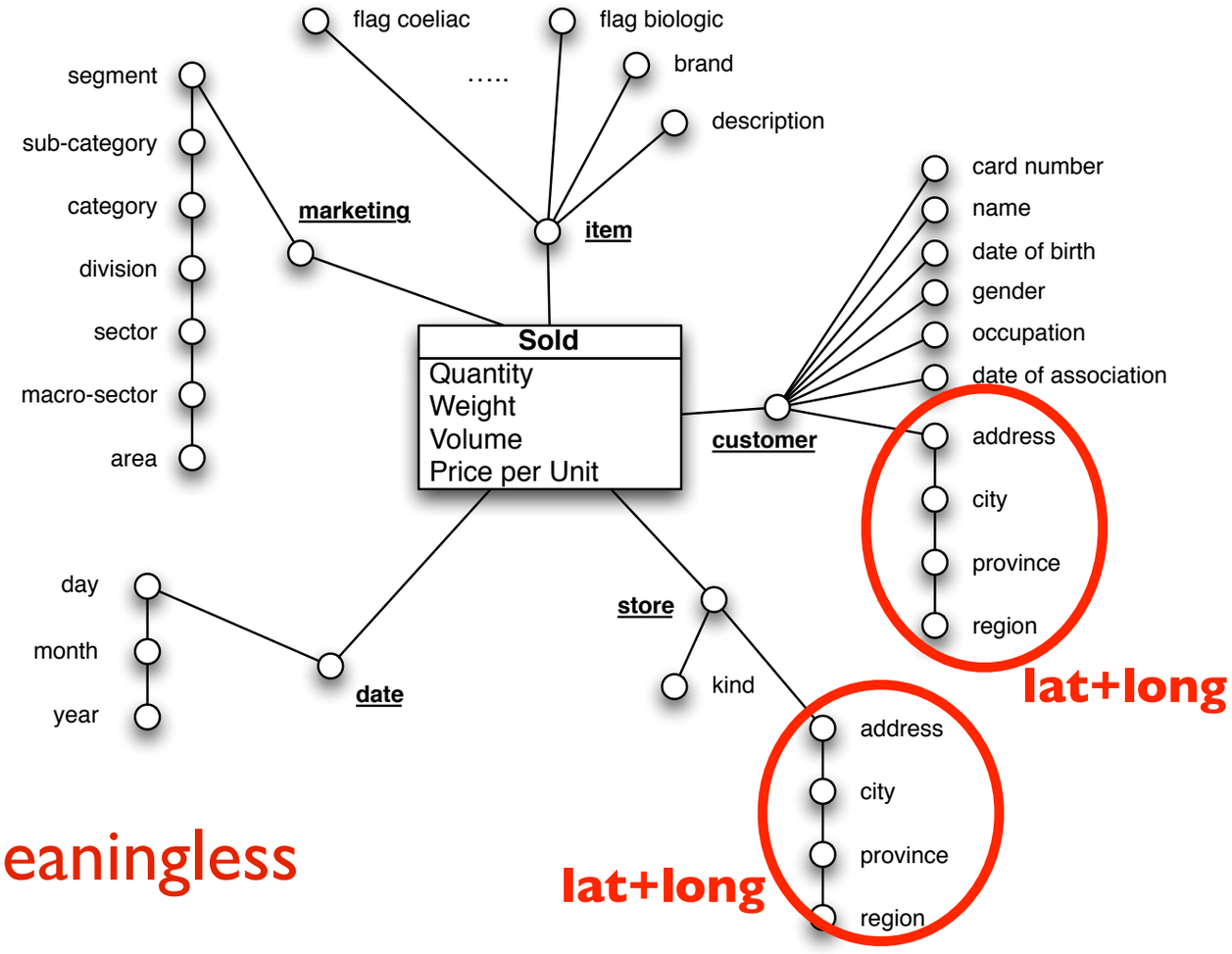


Data

- largest italian retail distribution Company
- time window: 2007-01-01 / 2012-12-31
- 1,066,020 active and recognizable customers
- 138 stores over the whole Italy's west coast
- 345,208 different items
- 2 Billions purchases

FILTER:

- one city (Leghorn)
- customers within 5km from a store
- segment, not item
- removed products too frequent and meaningless



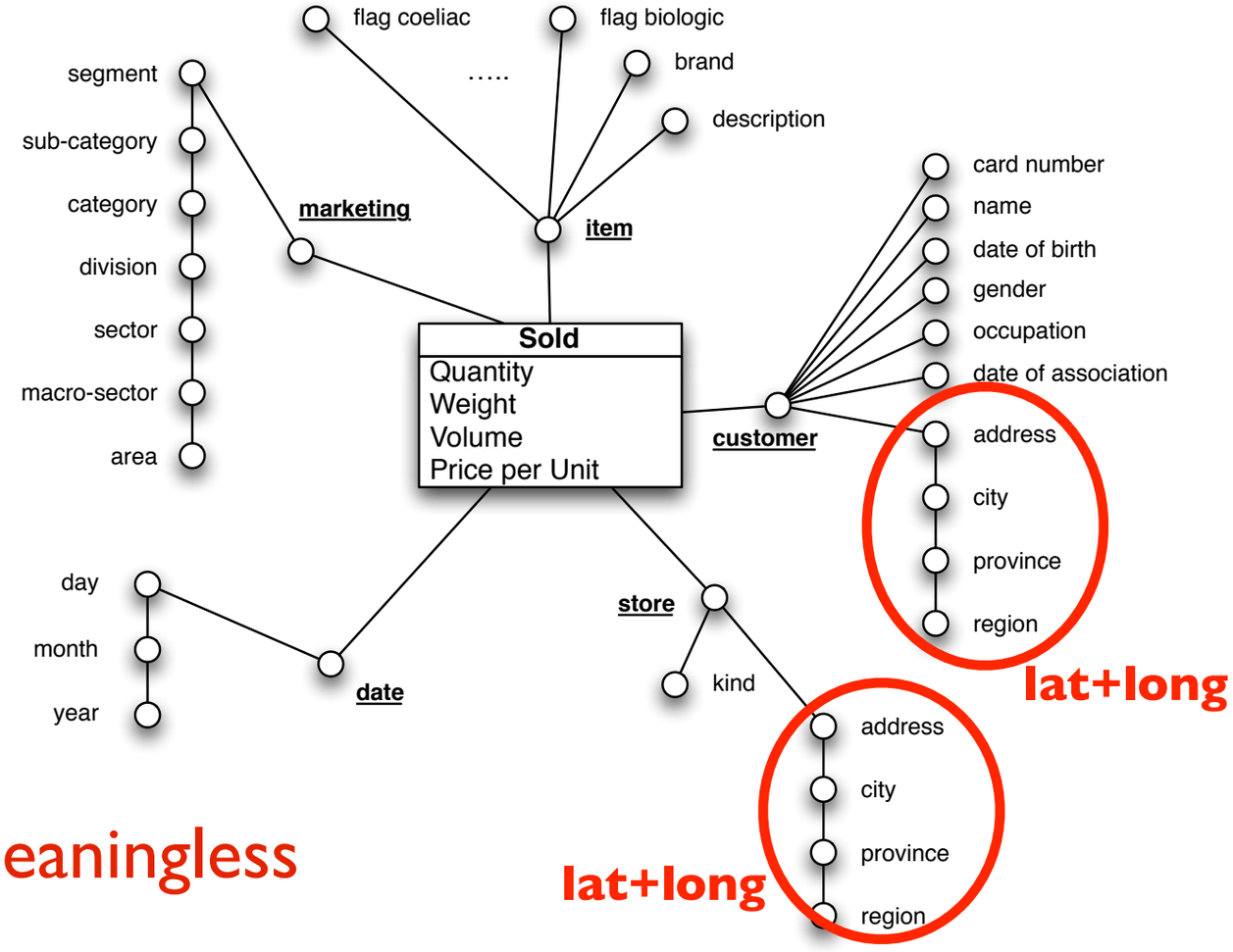
Data

After the filter:

- 60,366 customers
- 4,567 segments
- 107,371,973 total purchases

FILTER:

- one city (Leghorn)
- customers within 5km from a store
- segment, not item
- removed products too frequent and meaningless

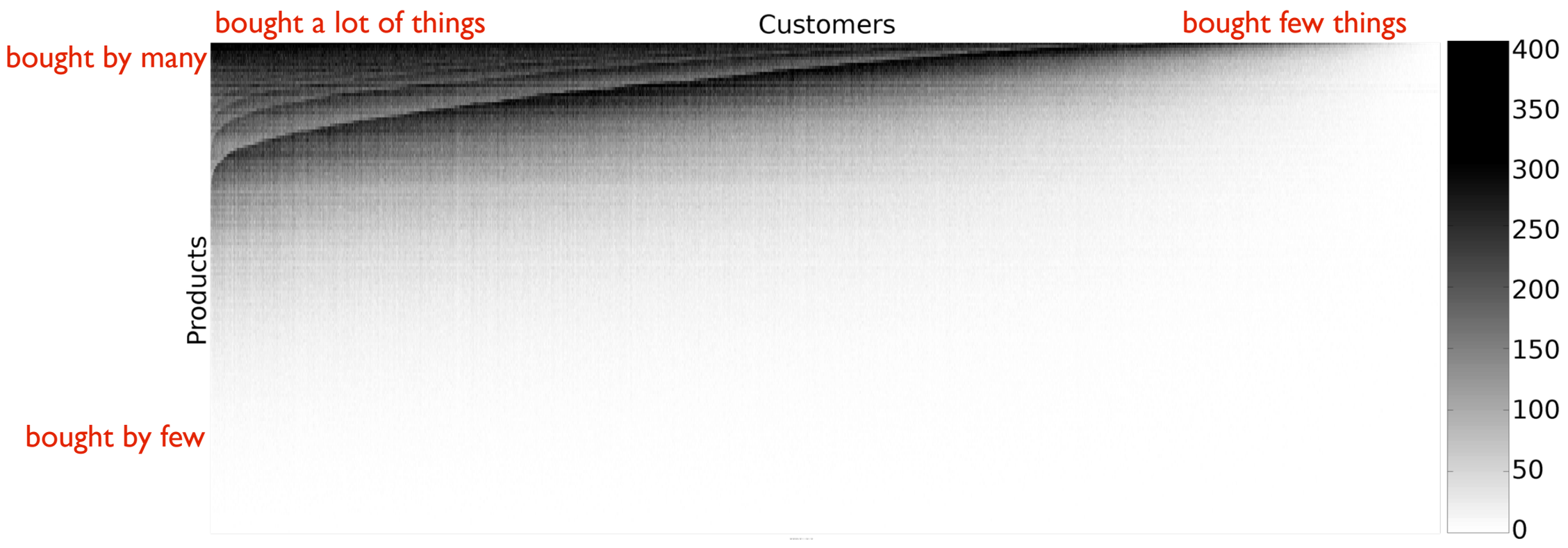


The data model

- Bipartite Network (customer-segment)
- not all purchases, just the ones *"meaningful" (lift)*
- *sort products and customers by quantities*

The data model

- Bipartite Network (customer-segment)
- not all purchases, just the ones *"meaningful" (lift)*
- *sort products and customers by quantities*



The product sophistication

- we calculate the sums of the purchase matrix for each product and customer
- we need to **correct these sums recursively**: we need to calculate the avg level of sophistication of the **customers**' needs by looking at the avg sophistication of the **products** that they buy, and then use it to **update** the avg sophistication of these products
- so, we take the eigenvector associated with the second largest eigenvalue (that is associated with the variance in the system) [HITS variation]

The product sophistication

less sophisticated

p_i	PS
Regular Bread	0.252
Red Meat	0.266
Artichokes	0.275
Pasta	0.275
Rabbit Meat	0.278

more sophisticated

p_i	PS
Winter Suit 3-12yo	0.796
TV 29”	0.769
DVD Readers	0.754
Hair Spray	0.742
8mm Cameras	0.739

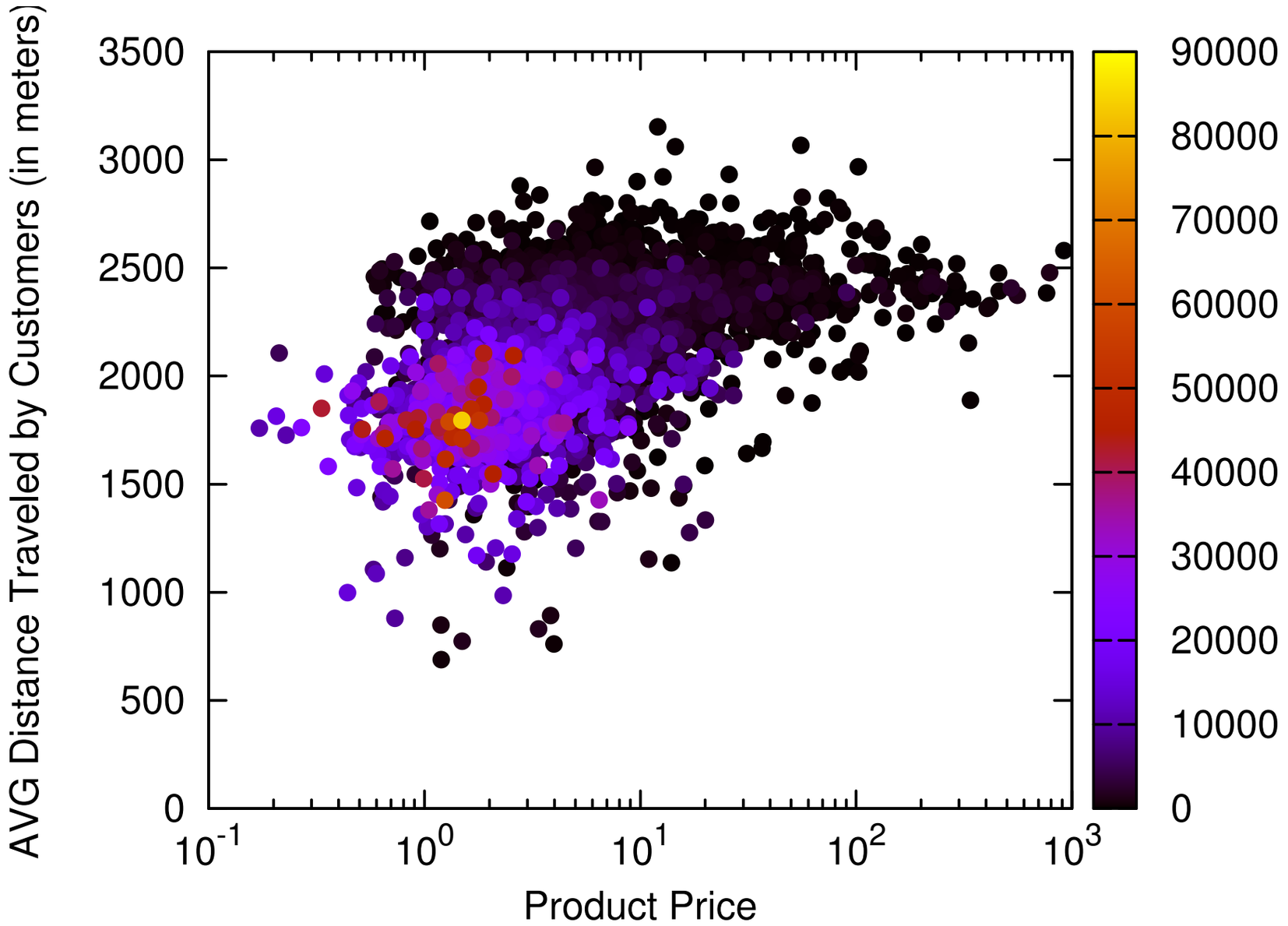
...and what about distance to travel?

- Intuitively...
  - higher prices mean longer distances to travel
  - higher quantities mean shorter distances to travel
  - higher sophistication means longer distances to travel

distance VS prices

- each dot is a purchase representative
- if a customer bought products of the same price in different shops, than the distance is weighted with the price

$$d(c_i, p_j) = \sum_{\forall s \in S} \frac{p_j(c_i, s) \times d(c_i, s)}{p_j(c_i, *)}$$

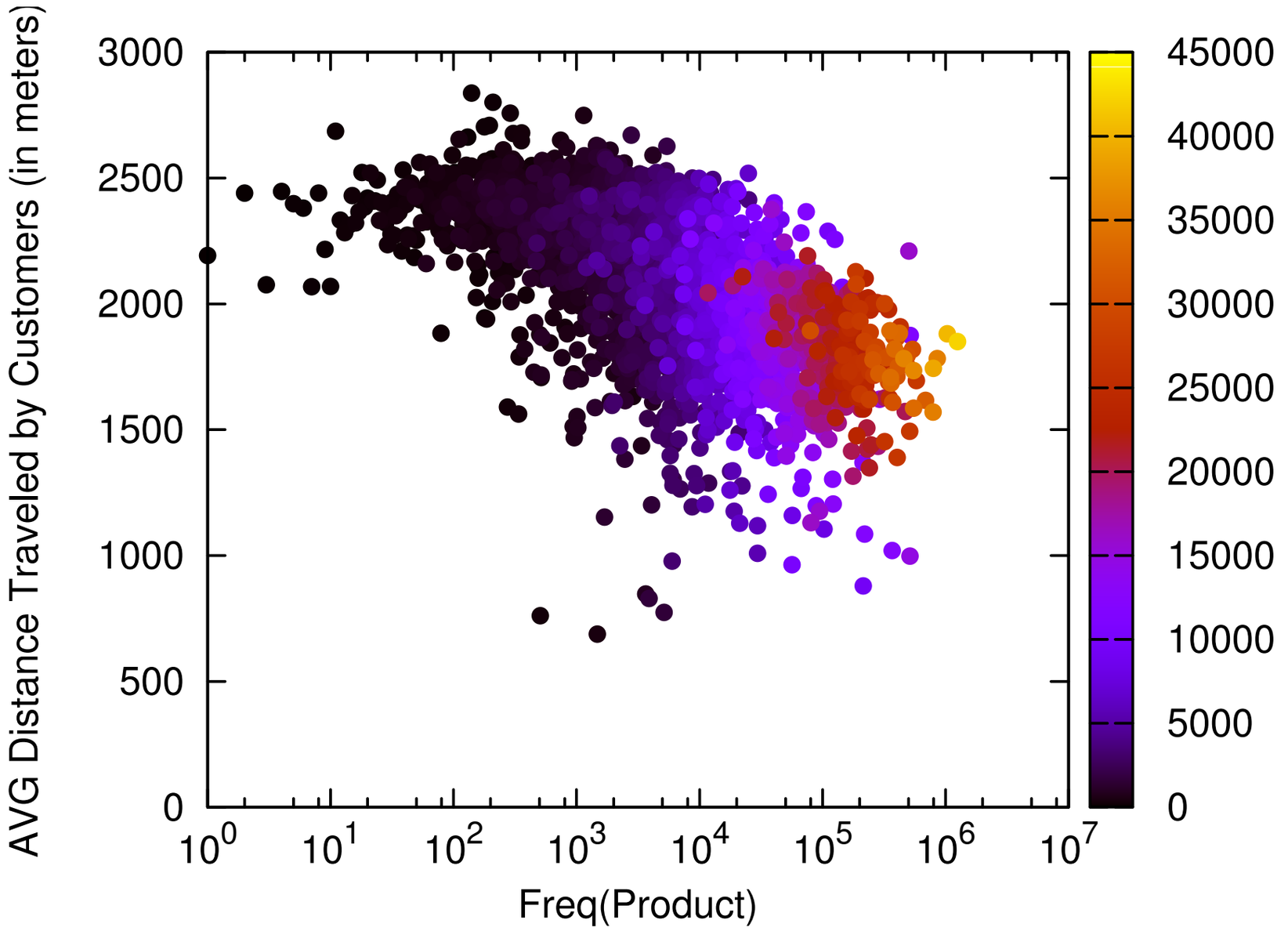


log-normal regression f(x) = a log x + b $R^2 = 17.25\%$

Santa Clara, October 7th 2013 - Diego Pennacchioli, National Research Council of Italy (Pisa) - diego.pennacchioli@isti.cnr.it

distance VS "popularity"

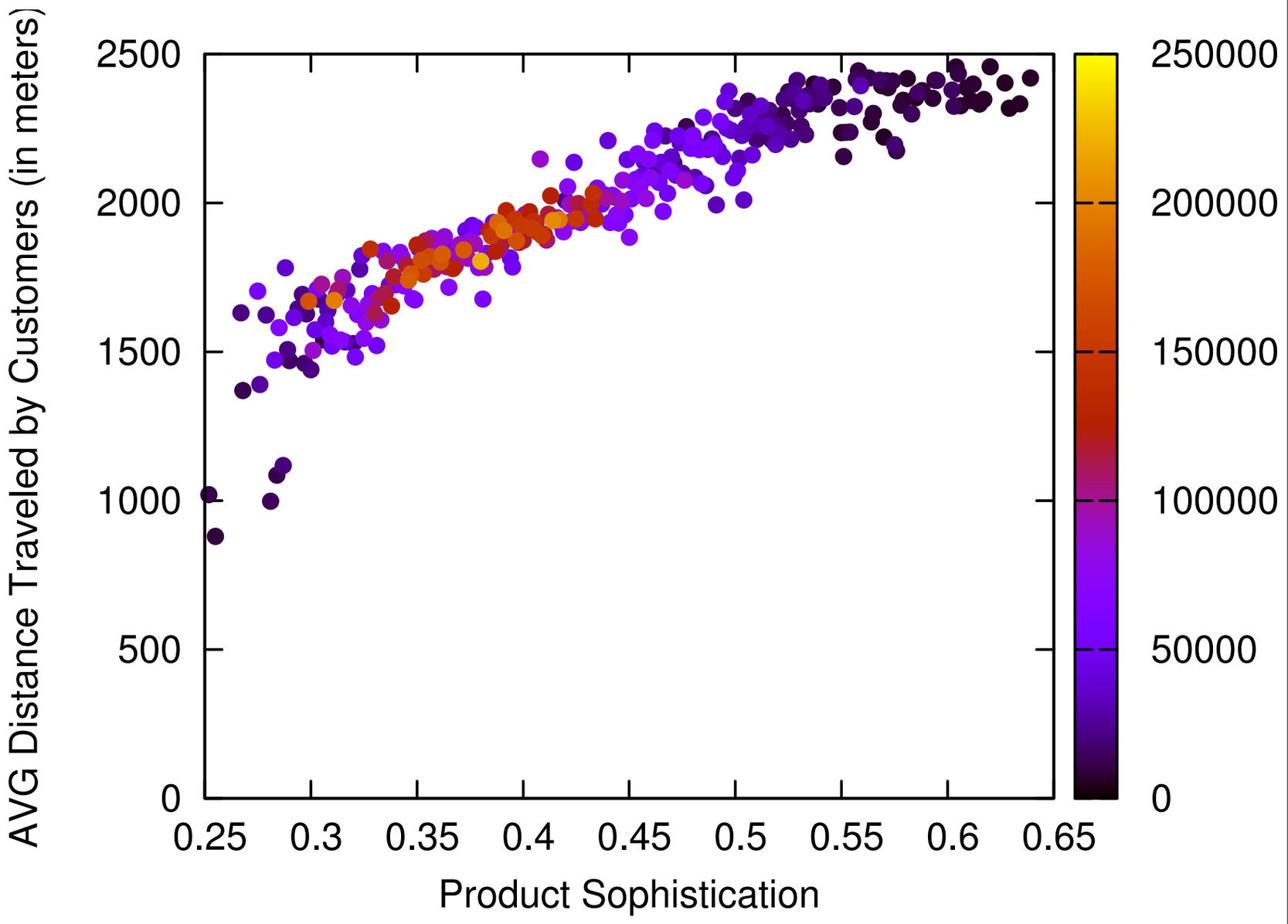
- each dot is a purchase representative
- if a customer bought products of the same frequency in different shops, than the distance is weighted with the frequency



log-normal regression $f(x) = a \log x + b$ $R^2 = 32.38\%$

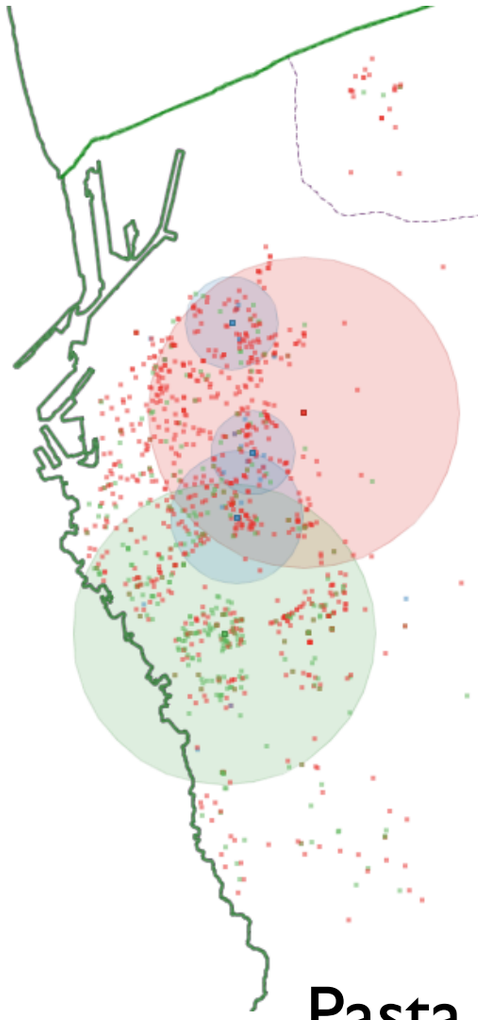
distance VS sophistication

- each dot is a purchase representative
- if a customer bought products of the same sophistication in different shops, than the distance is weighted with the sophistication

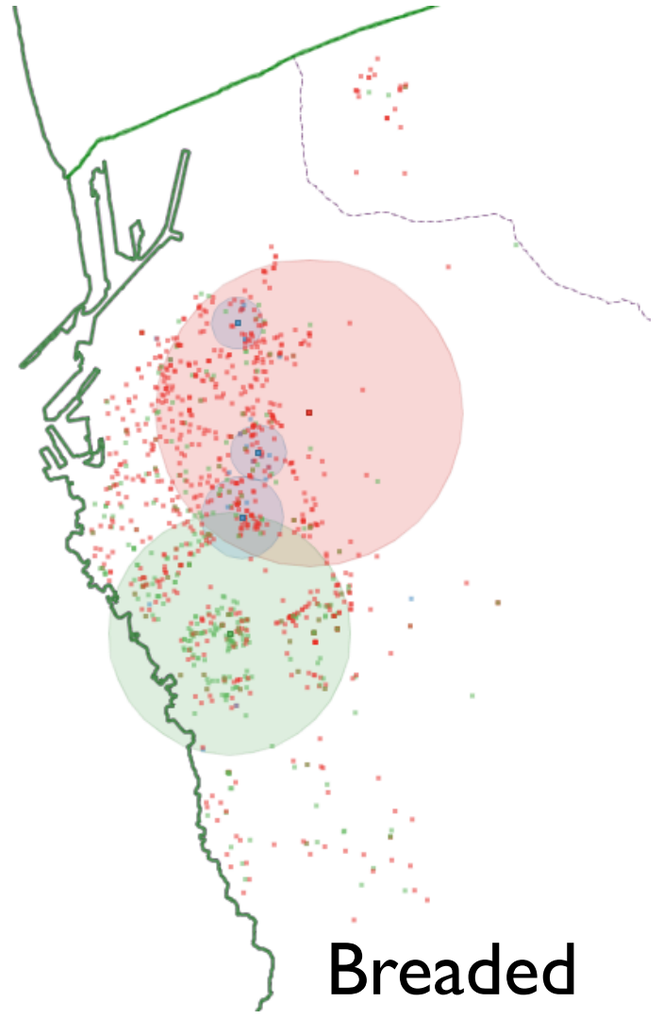


log-normal regression $f(x) = a \log x + b$ $R^2 = 85.72\%$

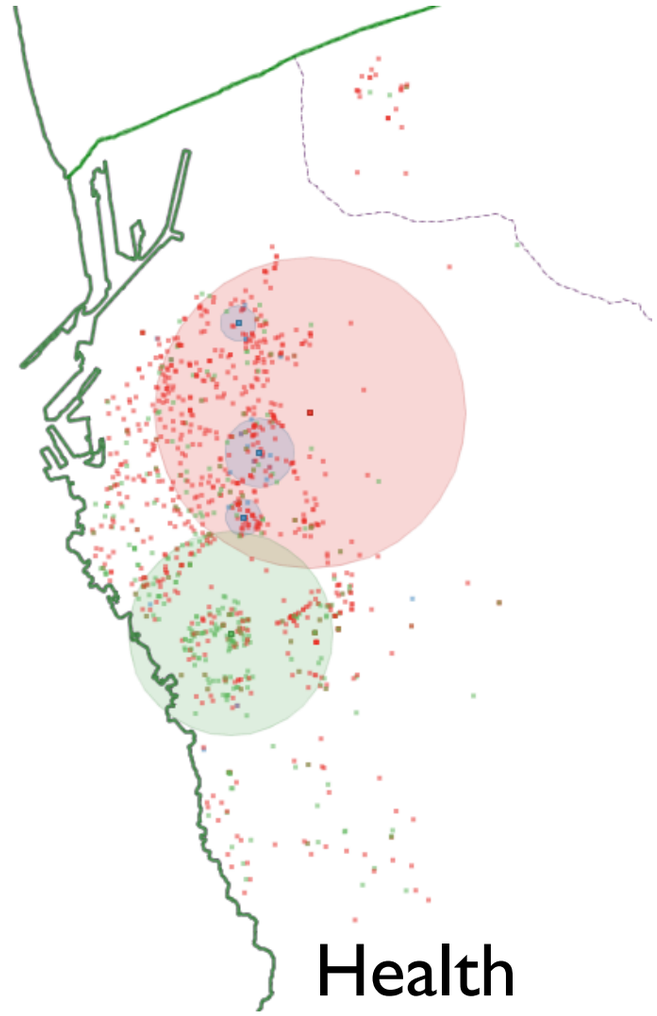
the power of shops



Pasta



Breaded
Frozen Fish

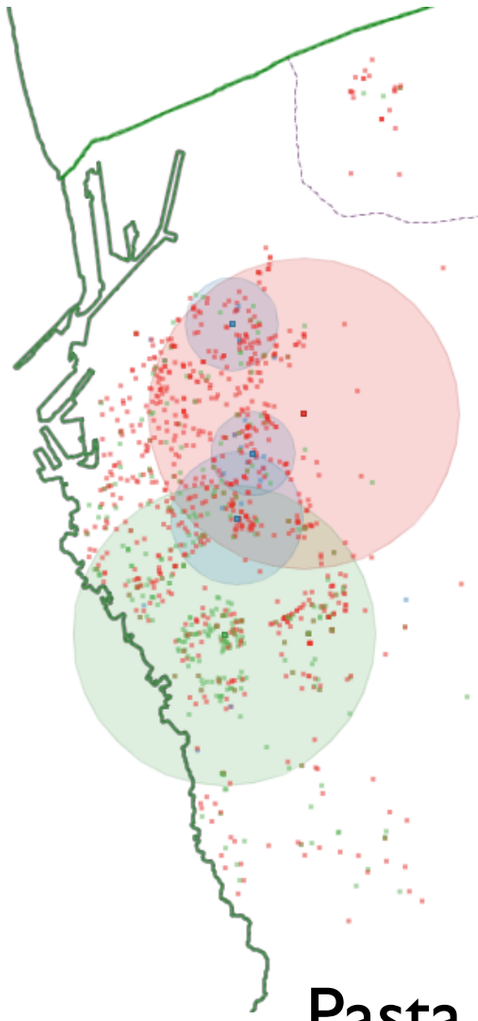


Health
Testers

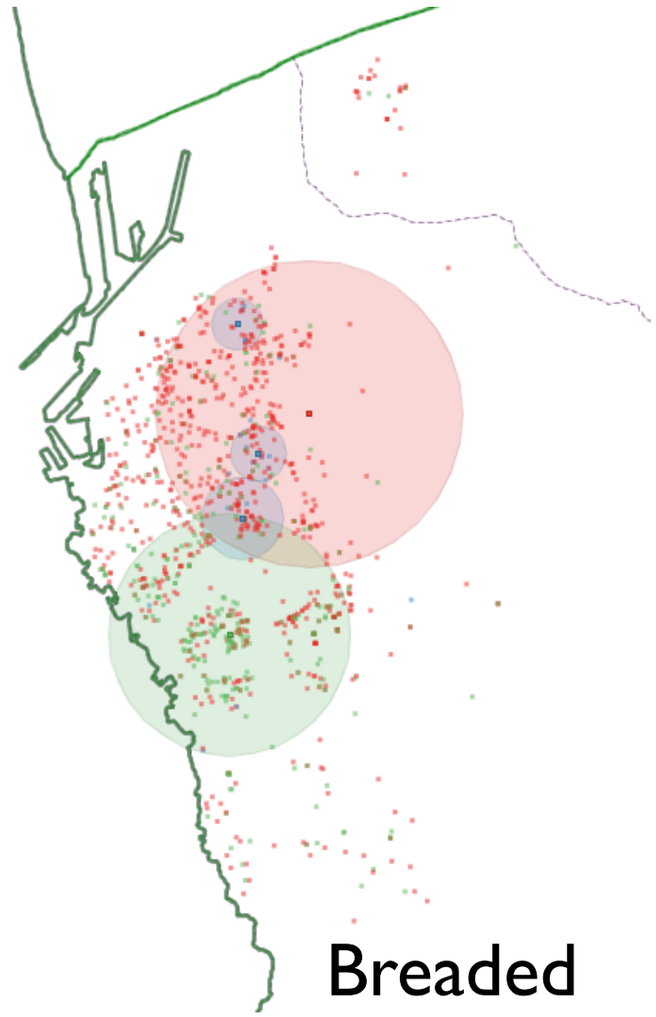
- Iper
- Super
- Gestin

Santa Clara, October 7th 2013 - Diego Pennacchioli, National Research Council of Italy (Pisa) - diego.pennacchioli@isti.cnr.it

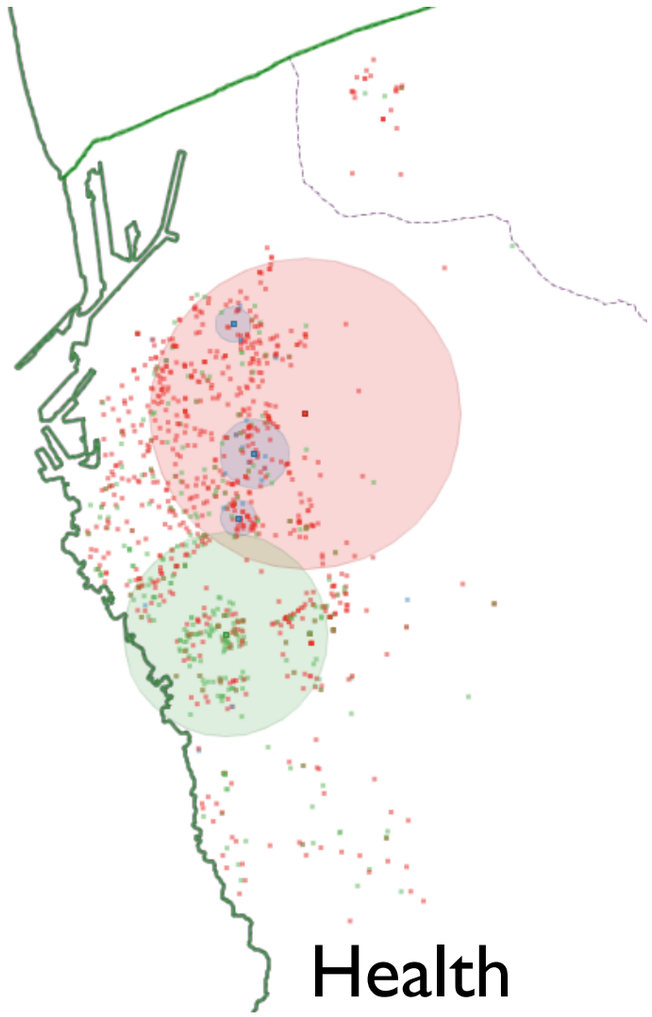
the power of shops



Pasta



Breaded
Frozen Fish



Health
Testers

- Iper
- Super
- Gestin

Shop Type	AVG PS	AVG Distance
Iper	0.49	2,392
Super	0.46	1,721
Gestin	0.43	869

Conclusions and Future Work

- introduced the concept of "range effect"
- shown the interplay with prices (low), popularity (medium) and sophistication (high)
- ...and if we use the distance as proxy for desire

- distance "as the crow flies" vs viability
- what about classification?
- ...and more complex measures?

Thank you for your attention!

Questions?

The product sophistication

- this formulation is very sensitive to noise (products bought by a very narrow set of customers)
  - Three-step strategy:
    - calculate eigenvectos of a restricted number of popular products
    - we use the estimate of the sophistication of these products to estimate the sophistication of all customers
    - we use the estimated sophistication of customers to have the final sophistication of the entire set of products
- at the end, we normalize $$PS(p) = \frac{\vec{K}(p) - \min(\vec{K})}{\max(\vec{K}) - \min(\vec{K})}$$