furio.camillo@unibo.it

OSSERVATORIO FEDELTÀ

Making Loyalty with Big Data: What challenges?

(Fare Loyalty con i Big Data: quali sfide?)

DataScienceLAB
is Bologna

furio camillo

Data Science LAB

Department of Statistical Sciences
Alma Mater Studiorum
University di Bologna - Italy
furio.camillo@unibo.it

Concepts cloud

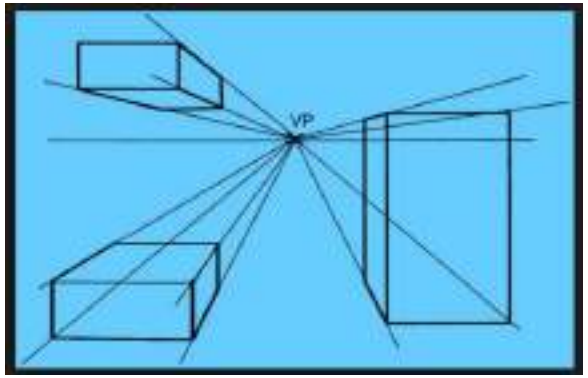
Data Science LAB

furio.camillo@unibo.it

Big-data, self-selection, predictions in real-time, intangible variables, opinion and sentiment manipulation: the landing in the future of business statisticians

The basic theme is to share and to discuss some recent experiences relating to the application of statistical analysis in companies and organizations, taking into account the scenario of context in order to clarify what are the frontiers of the future and the challenges of the present

New problems vs classical problems
New solutions vs classical solutions



Points of view:
Business
Data Analyst
Statistician

Context:
loyalty

Concepts cloud

furio.camillo@unibo.it

- ***Big data*** is a term for data sets that are so large or complex that traditional data processing applications are inadequate to deal with them. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy
- The term "big data" often refers simply to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set
- Accuracy in big data may lead to more confident decision-making, and better decisions can result in greater operational efficiency, cost reduction and reduced risk

furio.camillo@unibo.it

- Volume: big data doesn't sample; it just observes and tracks what happens
- Velocity: big data is often available in real-time
- Variety: big data draws from text, images, audio, video; plus it completes missing pieces through data fusion
- Machine Learning: big data often doesn't ask why and simply detects patterns
- Digital footprint: big data is often a cost-free by-product of digital interaction
- The growing maturity of the concept more starkly delineates the difference between big data and Business Intelligence

furio.camillo@unibo.it

- Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends, etc..
- Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships and dependencies, or to perform predictions of outcomes and behaviors.

Visualize data to decide: a dream

furio.camillo@unibo.it

TIBCO Spotfire

tableau

QlikView

IBM

IBM Watson Analytics

furio.camillo@unibo.it

The case against a paradigm shift in the way we use data

David Hand

Professor David Hand

A paradigm shift is a fundamental change in the basic concepts and practices of a discipline. Thomas Kuhn, who introduced the phrase in the context of scientific advances, contrasted it with normal science, which he defined as 'scientific work carried out within the context of an existing theory'[1]. So what might we mean by a paradigm shift in the way we

SUMMARY

- Although there have been advancements in the three dimensions of the data paradigm – data capture, data analysis and data storage – these are incremental developments, not fundamental changes in practice.

How do you respond when you hear the phrase 'big data'?

furio.camillo@unibo.it

How do you respond when you hear the phrase 'big data'?

David Hand


Professor David Hand

Probably a resigned sigh. 'Big data' is proclaimed as the answer to humanity's problems. However, while it's true that large data sets, a consequence of modern data capture technologies, do hold great promise for interesting and valuable advances, we should not fail to recognise that they also come with considerable technical challenges. The easiest of these lie in the data manipulation aspects of data science (the searching, sorting, and matching of large sets) while the toughest lie in the essentially statistical inferential aspects. The notion that one nowadays has 'all' of the data for any particular context is seldom true or relevant. And big data come with the data quality challenges of small data along with new challenges of its own.

Un sospiro rassegnato. 'Big Data' è proclamata come la risposta ai problemi dell'umanità.
Mentre è vero che, grazie alle tecnologie moderne di data collection, grandi insieme di dati hanno un grande potenziale per progressi interessanti, dobbiamo riconoscere che nello stesso tempo si presentano con notevole sfide tecniche. Le più semplici sono legate agli aspetti di data manipulation (l'ordinamento, il search ed il merge di grandi insiemi dei dati) mentre gli aspetti più difficili sono quelli inferenziali.
L'idea che oggi si disponga di 'tutti' i dati per ogni contesto particolare è raramente vera o rilevante. Oltre ai problemi di qualità di 'Small data', i 'Big data' presentano quindi delle sfide proprie e peculiari.

furio.camillo@unibo.it

Dealing with data generated by non-experimental studies

- Big data
- Algorithms vs. Statistical approach
- Professional skills

- **BIG-Propensity score FOR BIG-Data!!!**
- Cookies and digital (and real) footprint as data source

- No-structured data use (opinions)

furio.camillo@unibo.it

BIG DATA & SOCIETY

Commentary

Big Data & Society

Big Data and the danger of being precisely inaccurate

Daniel A McFarland and H Richard McFarland

$$S^2 = \frac{\sum_i (Xi - \bar{X})^2}{n}$$

Abstract

Social scientists and data analysts are increasingly making use of Big Data in their analyses. These data sets are often "found data" arising from purely observational sources rather than data derived under strict rules of a statistically designed experiment. However, since these large data sets easily meet the sample size requirements of most statistical procedures, they give analysts a false sense of security as they proceed to focus on employing traditional statistical methods. We explain how most analyses performed on Big Data today lead to "precisely inaccurate" results that hide biases in the data but are easily overlooked due to the enhanced significance of the results created by the data size. Before any analyses are performed on large data sets, we recommend employing a simple data segmentation technique to control for some major components of observational data biases. These segments will help to improve the accuracy of the results.

Keywords

Big Data, bias, segmentation, sociology, statistics, inaccuracy

Data Science
LAB

furio.camillo@unibo.it

Big data: are we making a big mistake?

By Tim Harford

Big data is a vague term for a massive phenomenon that has rapidly become an obs
entrepreneurs, scientists, governments and the media

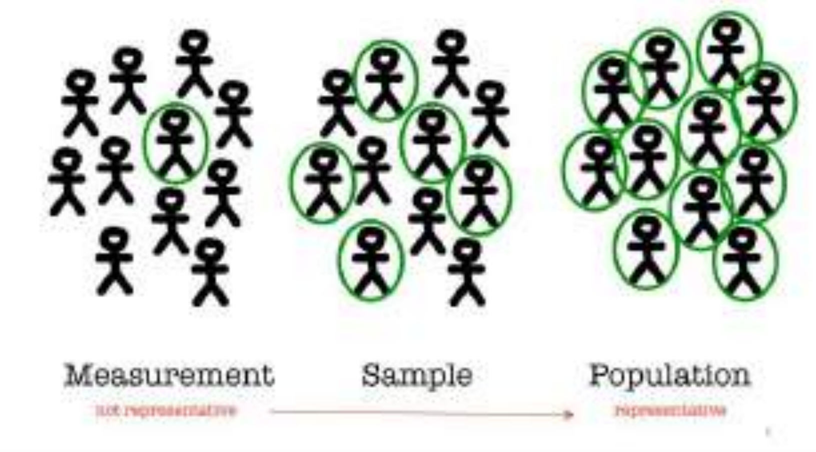
Professor Viktor Mayer-Schönberger of Oxford's Internet Institute, co-author o
data set is one where "N = All" – where we no longer have to sample, but we hav
not estimate an election result with a representative tally: they count the votes –
of sampling bias because the sample includes everyone.

But is "N = All" really a good description of most of the found data sets we are considering? Probably not. "I would challenge the notion that one could ever have all the data," says Patrick Wolfe, a computer scientist and professor of statistics at University College London.

An example is Twitter. It is in principle possible to record and analyse every message on Twitter and use it to draw conclusions about the public mood. (In practice, most researchers use a subset of that vast "fire hose" of data.) But while we can look at all the tweets, Twitter users are not representative of the population as a whole. (According to the Pew Research Internet Project, in 2013, US-based Twitter users were disproportionately young, urban or suburban, and black.)

furio.camillo@unibo.it

Dealing with data generated by non-experimental studies (sometimes with really self-selected samples)



Estimating Causal Effects in Observational Studies Using Electronic Health Data: Challenges and (some) Solutions

Elizabeth A. Stuart
Johns Hopkins Bloomberg School of Public Health, estuart@jhsph.edu

Eva DuGoff
Johns Hopkins Bloomberg School of Public Health, dugoff@wisc.edu

Michael Abrams
The University of Maryland Baltimore County, mabrams@hilltop.umbc.edu

David Salkever
UMBC, salkever@umbc.edu

……. using Propensity Score approach and Counterfactual Frame (Rubin)

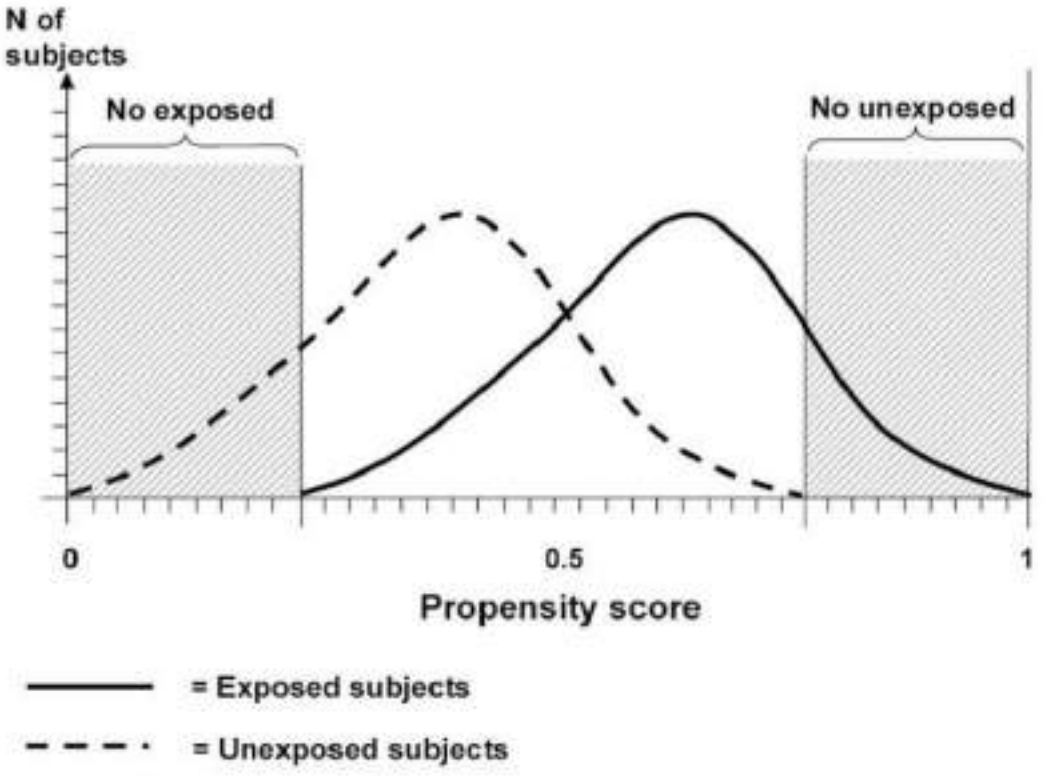
furio.camillo@unibo.it

UCL

Electronic Health Records versus Randomised Controlled Trials

	Electronic Health Records	Randon Trials
Data collection	Clinical sessions	At fixed
Data	Coded clinical records Read or ICD-10, measurements	Intervie questior measur
Missing data	Well people have less data	Randon
Size	Millions	From hu thousan
Treatment	Selective	Randon

Propensity scores



Data Science LAB

furio.camillo@unibo.it

Potential outcome framework

A causal effect is the comparison of the outcome that would be observed with the interventions (treatment) and without intervention, both measured at the same point in time (D. B. Rubin, R.P. Waterman, 2006)

Change in blood pressure (mm mercury)

Pill=yes Pill=no

subject	$Y_t(u)$	$Y_c(u)$	$Y_t(u) - Y_c(u)$
Joe		5	-10
Mary	-10		-5
Sally	0		-10
Bob		-5	-15

$$ATE = \Big((-10)+(-5)+(-10)+(-15)\Big) / 4 = -40/4 = -10$$

If the selection process depends on the same covariates conditioning «the result», the experiment will be biased

In BIG-DATA, the use of experiments, in general, is not controlled

Data Science LAB

Main ingredients: Cookie!!!!

furio.camillo@unibo.it

```
HTTP/1.1 200 OK
Cache-Control: private
Content-Type: text/html
Set-Cookie: PREF=ID=5e66ffd215b4c5e6:
TM=1147099841:LM=1147099841:S=Of69MpW
Bs23xeSv0; expires=Sun, 17-Jan-2038 1
9:14:07 GMT; path=/; domain=.google.c
om
```

A **cookie** is a small piece of data sent from a website and stored in a user's web browser while the user is browsing that website. Every time the user loads the website, the browser sends the cookie back to the server to notify the website of the user's previous activity.

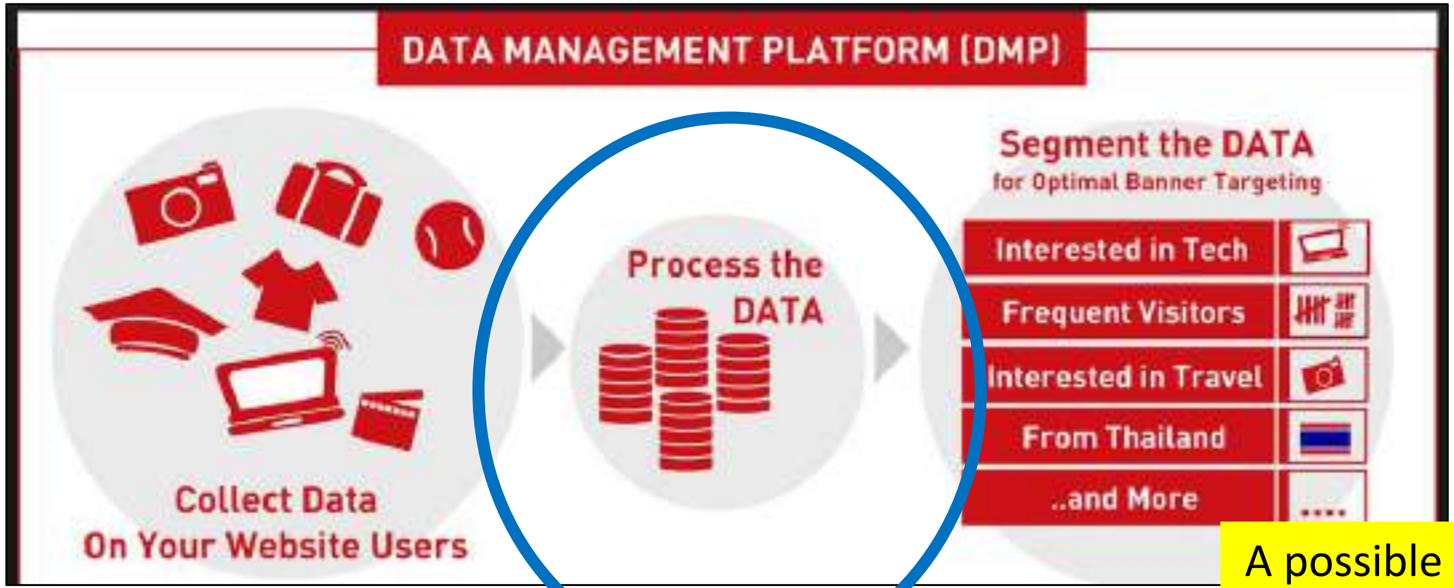
Track of a behaviour

Big-data, complex data, unstructured data: anyway, behavioural data *(in Italy more than 55million of tracked-cookies)*

Cookies are primarily used to provide services to the e-user (*especially in the mobile-apps*). In detail, main services from the use of cookies are:

- **fill the cart in commercial websites**
- **to allow login to a user**
- **customize the website according to user preferences**
- **management of a website: understand navigation, adjusting the spacing to web browsing needs and eliminate dead ends**
- **track paths of users: advertising companies should use such information to plan well communication and adjust in real time the messages according to the user's profile**

furio.camillo@unibo.it



what kind of modeling?
what kind of response?
what kind of prediction?
what kind of variables profiling?

A possible frame: supervised classification

- The a *posteriori* probability of a sample

$$P(Y=i\mid X)=\frac{p(X\mid Y=i)P(Y=i)}{p(X)}=\frac{\pi_i p_i(X)}{\sum_i \pi_i p_i(X)}=\underline{q_i(X)}$$

- Bayes Test:

$$q_1(x) \gtrless q_2(x) \Rightarrow \pi_1 p_1(x) \lessgtr \pi_2 p_2(x) \qquad \frac{p_1(x)}{p_2(x)} \gtrless \frac{\pi_2}{\pi_1}$$

- Likelihood Ratio:

$$\ell(X)=\frac{p_1(x)}{p_2(x)}$$

- Discriminant function:

$$h(X)=\ln \ell(x) = \ln p_1(x) - \ln p_2(x) \gtrless \underline{\ln \frac{\pi_2}{\pi_1}} \quad =0$$

furio.camillo@unibo.it

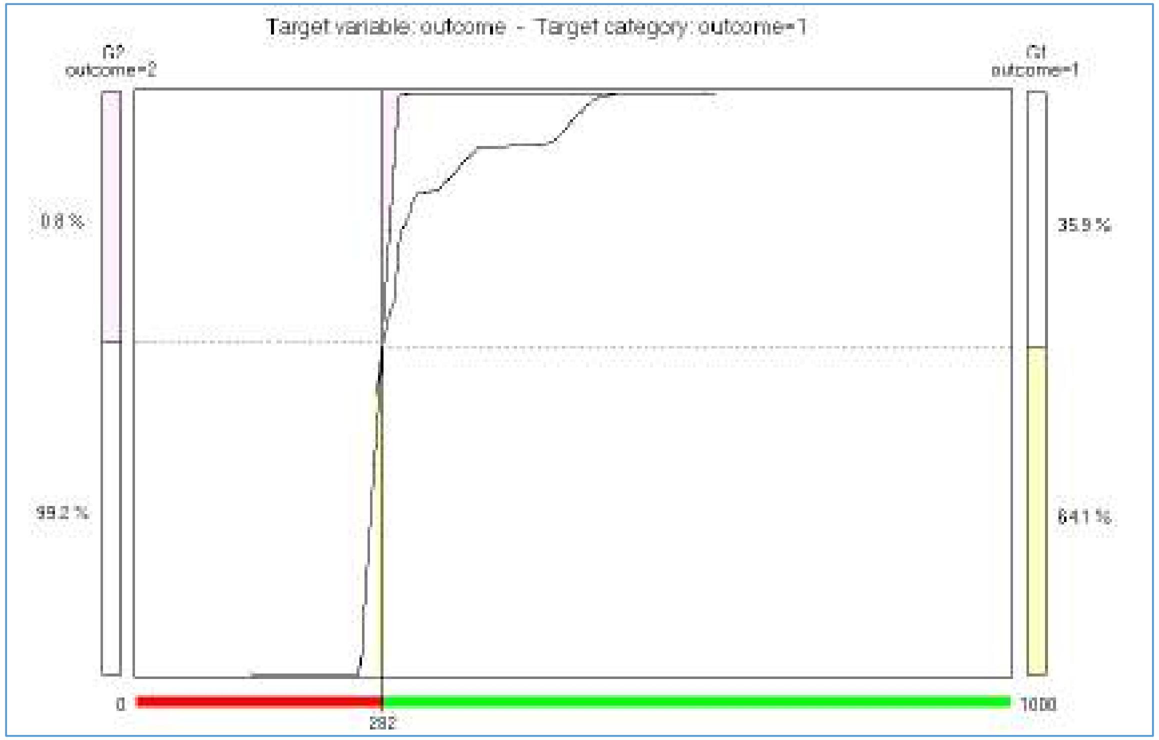
Predictive discriminant model

SUCCESSFUL EVENT: to complete the purchase of the new bank account (1=success; 2=no-success)

```
SELECTION OF CASES AN
ACTIVE CATEGORICAL VA
     5 VARIABLES
----------------------------------------------------------------------------------------
   3 . day                                                         (    7 CATEGORIES )
   4 . fascia_time                                                 (    6 CATEGORIES )
   5 . format2                                                     (   14 CATEGORIES )
   8 . site_name2                                                  (   48 CATEGORIES )
   9 . Campaign_name2                                              (   12 CATEGORIES )
----------------------------------------------------------------------------------------
```

Day
Time-band
Banner-format
Origin-site
Campaign-name

+ unsctructured infos about «user agent» string (coordinates of a Textual Binary Correspondence analysis)



Well predicted rate

Outcome=no-success: 99.2%
Outcome=success: 74.1%

$$s = \sum_{j=1}^{r} u_j X a^j = X \underbrace{\sum_{j=1}^{r} u_j a^j}_{scorecard}$$

Saporta scoring trasformation of parameters

furio.camillo@unibo.it

SITE NAME	score
Borsaltaliana(Websystem)	[illegible]
Confrontaconti(Confrontaconti)	315.52
T2O Media(T2O Media)	[illegible]
Ilsole24Ore(Websystem)	252.62
NULL	238.66
yahoo.it(Yahoo! Advertising So	221.83
Clickpoint Display(Clickpoint)	200.48
Facile.it(Facile.it)	198.48
Google Display Network(Google	170.80
Mutuosupermarket(Mutuosupermar	166.74
Performachine(Performachine)	155.51
Msn.it(Microsoft)	152.07
MilanoFinanza(Class)	148.95
Google SEM(Search Engine)	142.37
Arcus(Arcus)	142.13
Bing/Yahoo SEM(Search Engine)	139.01
MioJob(Manzoni Advertising)	138.16
Leonardo.it(Leonardo Adv)	134.70
Payclick(Payclick)	134.31
Criteo(Criteo IT)	133.15
RocketFuel(RocketFuel)	132.77
IlGiornaleOff.it(Websystem)	132.59
Trovolavoro(RCS MediaGroup)	132.45
Accuen Network IT(Accuen IT)	132.13
Italiaonline(Italiaonline)	132.12
Tradedoubler(Tradedoubler)	131.83
Clickpoint DEM(Clickpoint)	131.65
RCS Network(RCS MediaGroup)	128.17
MSN(Microsoft)	127.56
Webperformance(Webperformance)	127.03
LinkedIn(Linkedin)	125.33
ValueDem(ValueDem)	121.72
Casa.it(Casa.it)	117.80
TgAdv(Tg adv)	117.13
Monster(Monster)	113.23
Veesible(Veesible)	108.62
T2O(T2O)	102.00
Affaritaliani(Websystem)	96.65
Bluerating(Bluerating)	91.99
Webperformance DEM(Webperforma	91.70
Juice(Leonardo-Juice)	88.00
Teradata(Teradata)	87.22
Facebook(Facebook)	84.77
Yahoo Stream Ads(Yahoo! Advert	82.25
Cliccalavoro(Antevenio)	78.10
Leonardo Network(Leonardo-Juic	64.07
Digitouch(Digitouch)	59.86
AdKaora(AdKaora)	0.00

CAMPAIGN NAME	score
MR[illegible]	142,66
n[illegible]	115,19
C[illegible]	61,07
[illegible]	31,43
[illegible]	18,40
[illegible]	12,32
[illegible]	9,04
[illegible]	8,95
[illegible]	8,66
[illegible]	8,62
FE[illegible]	8,48
MR[illegible]	0,00

FORMAT	score
format2=Yahoo	426.45
format2=250x250	257.37
format2=205x205	171.14
format2=336x280	139.27
format2=160x600	137.28
missing category	137.01
format2=728x90	136.85
format2=300x250	136.76
format2=120x600	136.42
Libero-Virgilio	136.32
format2=Corriere	130.28
format2=MSN	130.02
format2=468x60	77.30
format2=300x600	0.00

DAY	score
venerdì	1.29
mercoledì	[illegible]
giovedì	0.58
martedì	0.56
domenica	0.47
sabato	0.19
lunedì	0.00

TIME	score
14-18	1.11
12-14	0.77
18-21	0.60
8-12	0.30
21-01	0.28
01-8	0.00

Example (scoring-points):

ConfrontoConti	315
Campaign xxxx	142
Banner format: 250x250	257
Day: friday	1
Time: 14-18	1

Total points (score) 716

Combination with high probabilty of success

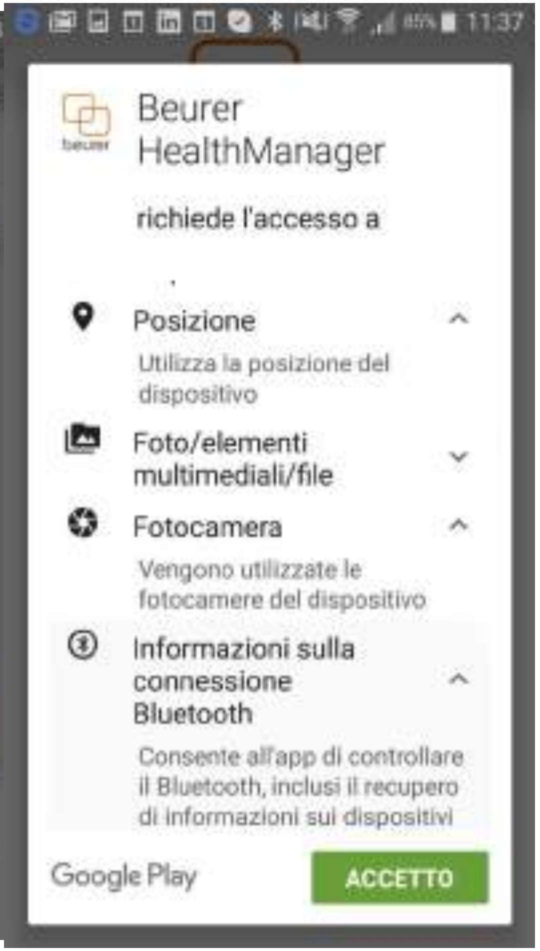
Using nonparametric models (for example knn) give better results because they use the non-linearity of relations. Next research: kernel discriminant analysis

PROBABILISTIC MACHINE LEARNING

furio.camillo@unibo.it

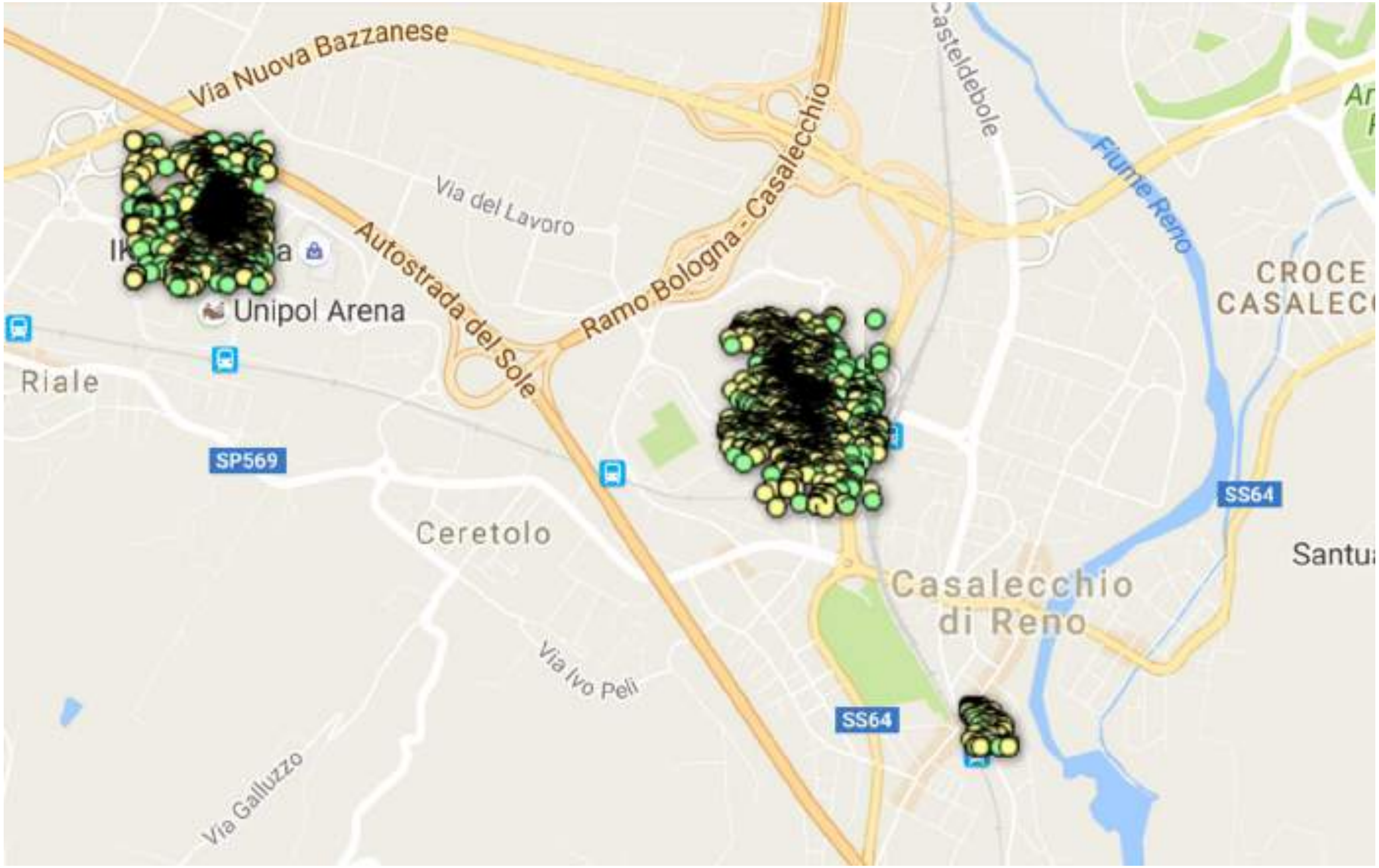
SITE NAME	score
BorsaItaliana(Websystem)	428.48
Confrontaconti(Confrontaconti)	315.52
T2O Media(T2O Media)	255.53
Ilsole24Ore(Websystem)	252.62
NULL	238.66
yahoo.it(Yahoo! Advertising So	221.83
Clickpoint Display(Clickpoint)	200.48
Facile.it(Facile.it)	198.48
Google Display Network(Google	170.80
Mutuosupermarket(Mutuosupermar	166.74
Performachine(Performachine)	155.51
Msn.it(Microsoft)	152.07
MilanoFinanza(Class)	148.95
Google SEM(Search Engine)	142.37
Arcus(Arcus)	142.13
Bing/Yahoo SEM(Search Engine)	139.01
MioJob(Manzoni Advertising)	138.16
Leonardo.it(Leonardo Adv)	134.70
Payclick(Payclick)	134.31
Criteo(Criteo IT)	133.15
Criteo(Criteo IT)	133.15
RocketFuel(RocketFuel)	132.77
IlGiornaleOff.it(Websystem)	132.59
Trovolavoro(RCS MediaGroup)	132.45
Accuen Network IT(Accuen IT)	132.13
Italiaonline(Italiaonline)	132.12
Tradedoubler(Tradedoubler)	131.83
Clickpoint DEM(Clickpoint)	131.65
RCS Network(RCS MediaGroup)	128.17
MSN(Microsoft)	127.56
Webperformance(Webperformance)	127.03
LinkedIn(Linkedin)	125.33
ValueDem(ValueDem)	121.72
Casa.it(Casa.it)	117.80
TgAdv(Tg adv)	117.13
Monster(Monster)	113.23
Veesible(Veesible)	108.62
T2O(T2O)	102.00
Affaritaliani(Websystem)	96.65
Bluerating(Bluerating)	91.99
Webperformance DEM(Webperforma	91.70
Juice(Leonardo-Juice)	88.00
Teradata(Teradata)	87.22
Facebook(Facebook)	84.77
Yahoo Stream Ads(Yahoo! Advert	82.25
Cliccalavoro(Antevenio)	78.10
Leonardo Network(Leonardo-Juic	64.07
Digitouch(Digitouch)	59.86
AdKaora(AdKaora)	0.00

Technology no-cookie but geo-referred (3/4G, GPS)



furio.camillo@unibo.it

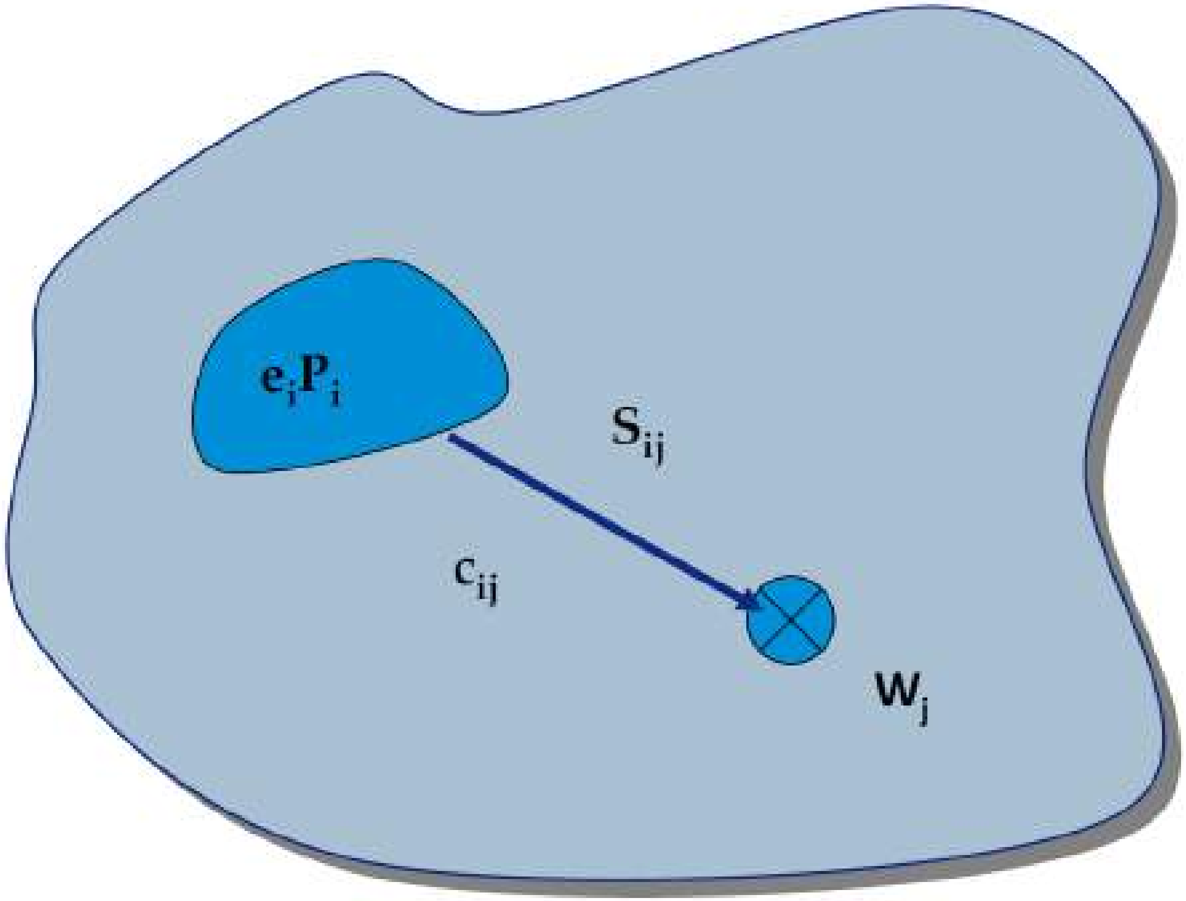
Example of smartphone track data



furio.camillo@unibo.it

The retail archetype

Spatial Interaction
Alan Wilson – UCL



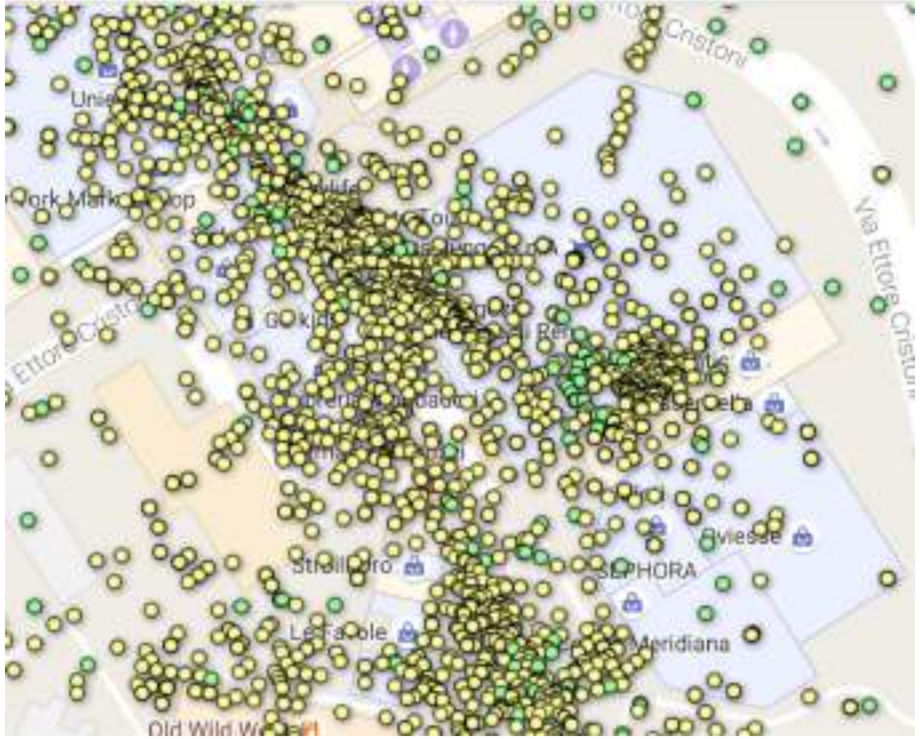
e_iP_i - demand in zone i

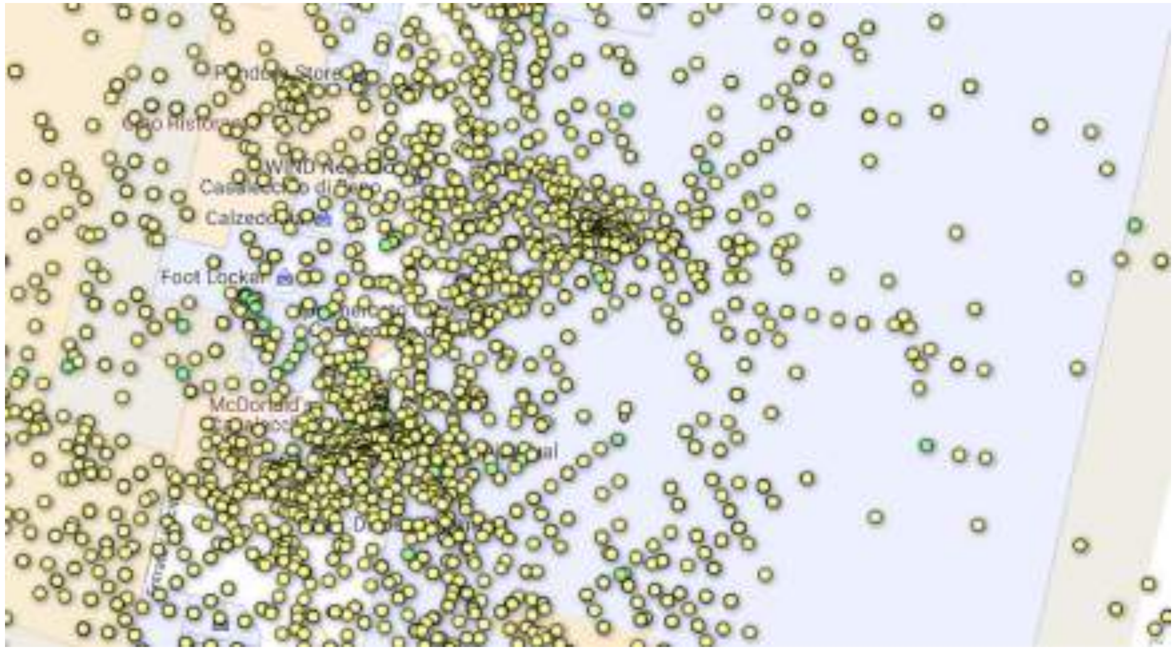
W_j - attractiveness of zone j

S_{ij} - flows between i and j

c_{ij} - cost of travel between i and j

furio.camillo@unibo.it





Resarch project about (with Beintoo)

1. **Modelling Spatial Interaction and natural Gravity areas**
2. **Indirect loyalty estimation**
3. **Churn timing**
4. **Budget time as a proxy for style profiling based on values or limbic propensities**

furio.camillo@unibo.it

Case: digital measurement of loyalty to drug (self-selection)

- USA, drug evaluation, 1.000.000 geo-tracked citizen (high social class)
- Evaluation: 2014 vs 2016
- Clustering using more than 150 geo-behavioural variables and 25 «hard» variables
- Global Imbalance Index (GI) to select only balanced clusters
- **Treatment**: year; **Outcome**: Indirect loyalty index about drugs use (high cholesterol, high blood pressure)
- **Collection**: by specific tracked IoS and Android APPs
- Qualitative information: open textual opinions

furio.camillo@unibo.it

2014 2016

You live in...	Percent	Percent	Chi-Square	P-Value
NEW YORK	30.79	40	4.27	0.64
LOS ANGELES	21.52	18.67		
CHICAGO	15.89	15		
DALLAS	9.6	8		
MIAMI	5.3	4		
SAN FRANCISCO	10.6	8.33		
HOUSTON	6.29	6		

Could you please tell me the total amount of your family gross annual earnings?	Percent	Percent	Chi-Square	P-Valu
Between 80,001 and 100.000 $	15.56	0	31.24	0.0
Between 100,001 and 125.000 $	19.54	18		
Between 125.001 and 150,000 $	20.53	23		
Between 150.001 and 170.000 $	15.89	13.67		
Between 170.001 and 200.000 $	11.59	14.67		
Between 200.001 and 250.000 $	9.6	14		
Over 250,000 $	7.28	16.67		

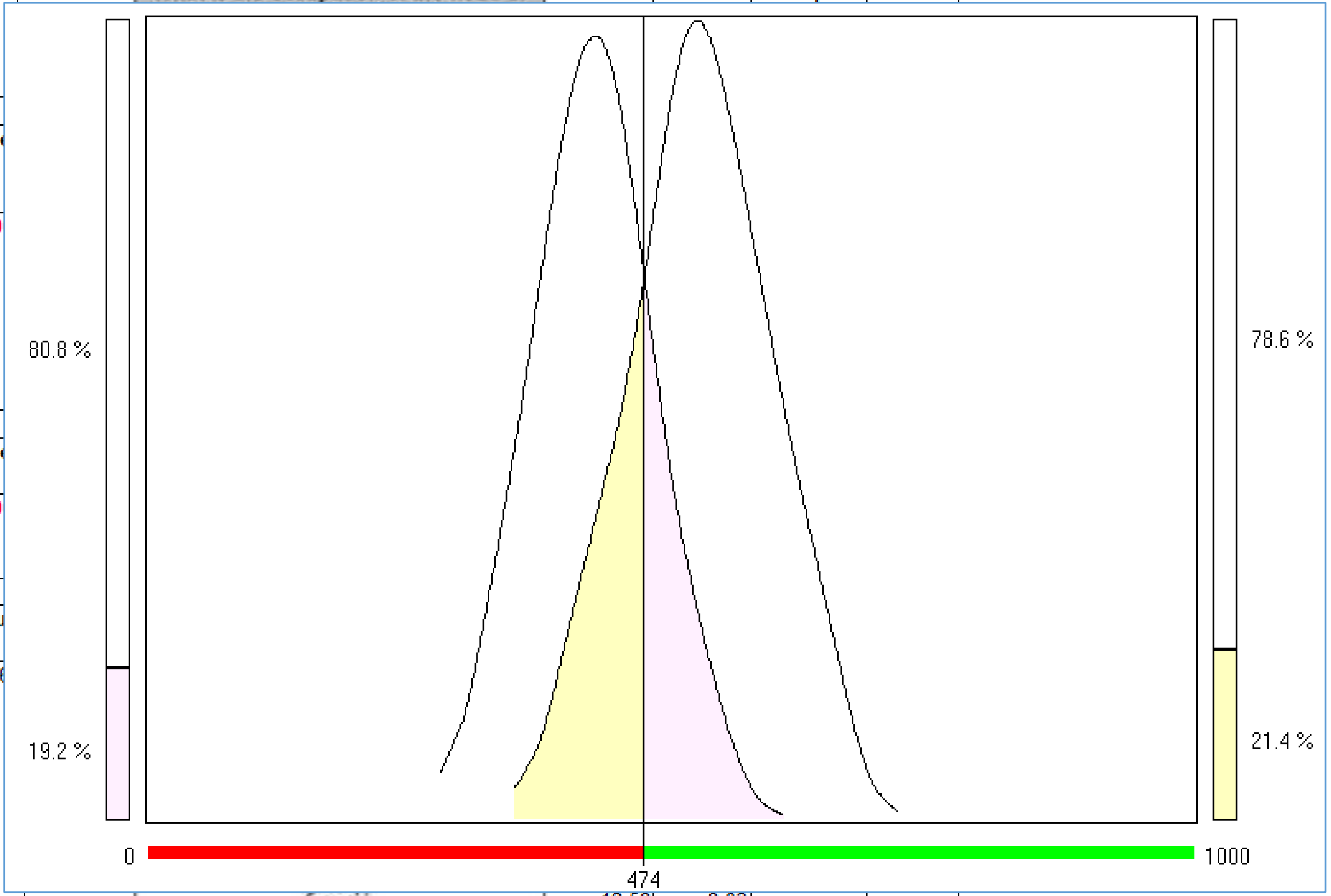
How old are you?	Percent	Percent	Chi-Square	P-Valu
25-35	38.08	26	21.91	0.0
36-45	31.46	22		
46-60	30.46	52		

Could you please tell me your occupation?	Percent	Percent	Chi-Square	P-Valu
Entrepreneur, manager, free-lance professional	20.86	20.67	4.74	0.
White collar	57.62	54.67		
Agent/self-employed	3.64	5		
Teacher/journalist	9.93	8		
Housewife	6.62	8.67		
Student	0.33	0.33		
Retired	0.66	2		
Unemployed	0.33	0.67		

2014 2016

Are you the head of the family, that is the chief income earner?	Percent	Percent	Chi-Square	P-Value
Yes	65.56	56.67	3.50	0.06
No	34.44	43.33		

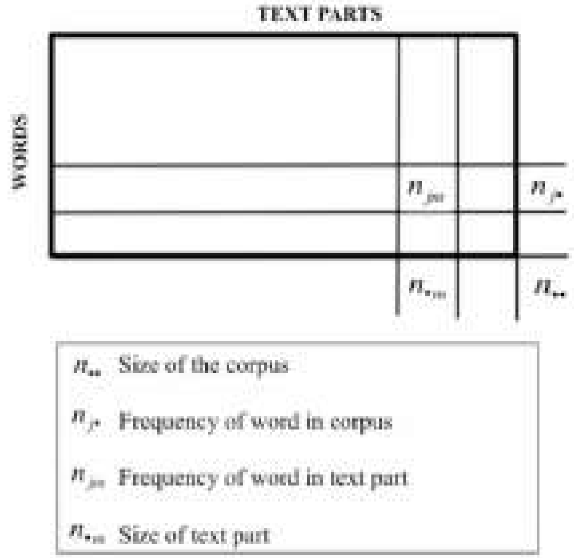
What is the occupation of the head of	Percent	Percent	Chi-Square	P-Value
5 o più	12.58	8.03		



furio.camillo@unibo.it

Specificity analysis (french approach, L.Lebart): negative opinions

Figure 1. Authors' elaboration from Lebart, Salem & Berry (1998) probabilistic scheme



1 For me, simvastatine was pure posion. I took it for four months and day by day felt more miserable. Enormous amount of bruising, scratching myself raw, recently muscle pain, could hardly walk any more. After stopping I felt drastically better.

2 Started off with rigid fingers and then muscle pain in my legs that was so bad I could hardly walk any more, stopped taking it 3 days ago and the muscle pain in my legs is near enough gone so that I can walk well again. Investigated for muscle degradation but nothing was detected. Next week back to the GP...

furio.camillo@unibo.it

Perspectives

- Interpretation of data vs. Big Data: complexity
- Scientific, rational, illuministic approach
- Y = f(X) (Correlation vs Causality)
- French school (1968): «*Le modèle est dans les données, mais il faut le chercher*»
- Navigation Hypothesis: Business KPIs
- Small data mixed with Big data
- Leadership in a Big-Data Project
- Technology does not necessarily reduce the number of people working in business

Data Science LAB

OSSERVATORIO FEDELTÀ

Making Loyalty with Big Data: What challenges?

(Fare Loyalty con i Big Data: quali sfide?)

furio camillo
Department of Statistical Sciences
Alma Mater Studiorum
University di Bologna - Italy
furio.camillo@unibo.it

In cooperation with Valentina Adorno

Data Science LAB

www.datasciencelab.it

A «Bologna city-brand» company

DataScienceLAB

is Bologna