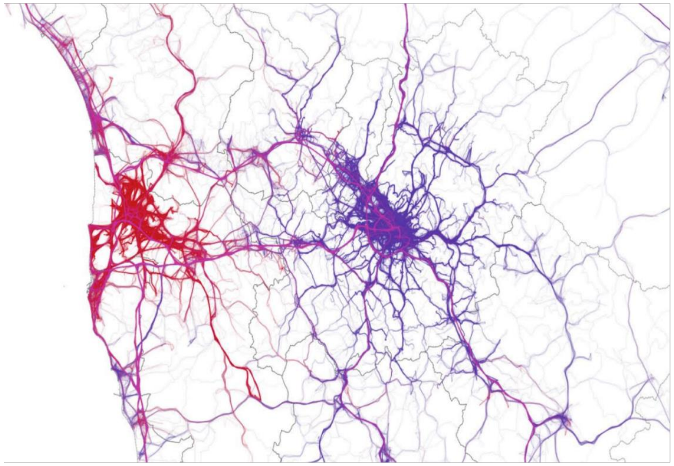
Ph.D. in Data Science



Data Science Colloquium

A Lunch Seminar Series

Session#1 (Fall 2017)

Data Science PhD

Monday, 27th November 2017, h.13:00

Dino Pedreschi, UNIPI: Data Science for pattern discovery and machine learning transparency

Wednesday, 6th December 2017. h 13:00

Tommaso Cucinotta, SSSA: Real-time cloud and big-data processing infrastructures

Wednesday, 13th December 2017, h.13:00

Marco Conti, IIT: From ego-networks to online social networks

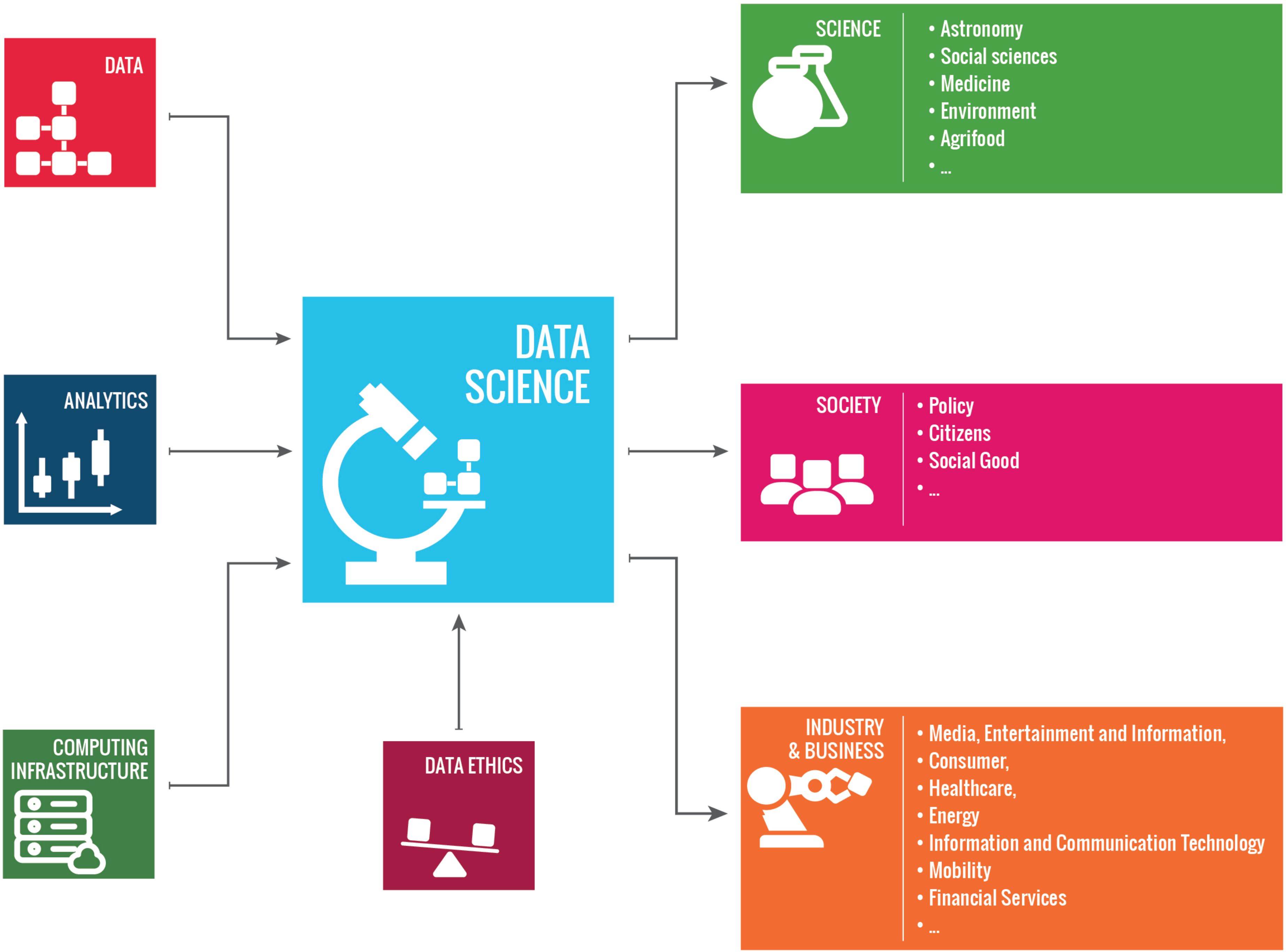
All seminars will take place at

AULA GERACE, Dipartimento di Informatica, Università di Pisa

Polo Fibonacci (ex Marzotto), Edificio C,

Largo Bruno Pontecorvo, 3 - PISA

Scuola Normale Superiore · Università di Pisa · Sant'Anna Scuola Universitaria Superiore Pisa · IMT School for Advanced Studies Lucca · Consiglio Nazionale delle Ricerche



DATA
ANALYTICS
COMPUTING INFRASTRUCTURE
DATA ETHICS
DATA SCIENCE
SCIENCE
• Astronomy
• Social sciences
• Medicine
• Environment
• Agrifood
• ...
SOCIETY
• Policy
• Citizens
• Social Good
• ...
INDUSTRY & BUSINESS
• Media, Entertainment and Information,
• Consumer,
• Healthcare,
• Energy
• Information and Communication Technology
• Mobility
• Financial Services
• ...

SoBigData Research Infrastructure

Data Science for Pattern Discovery and Machine Learning Transparency

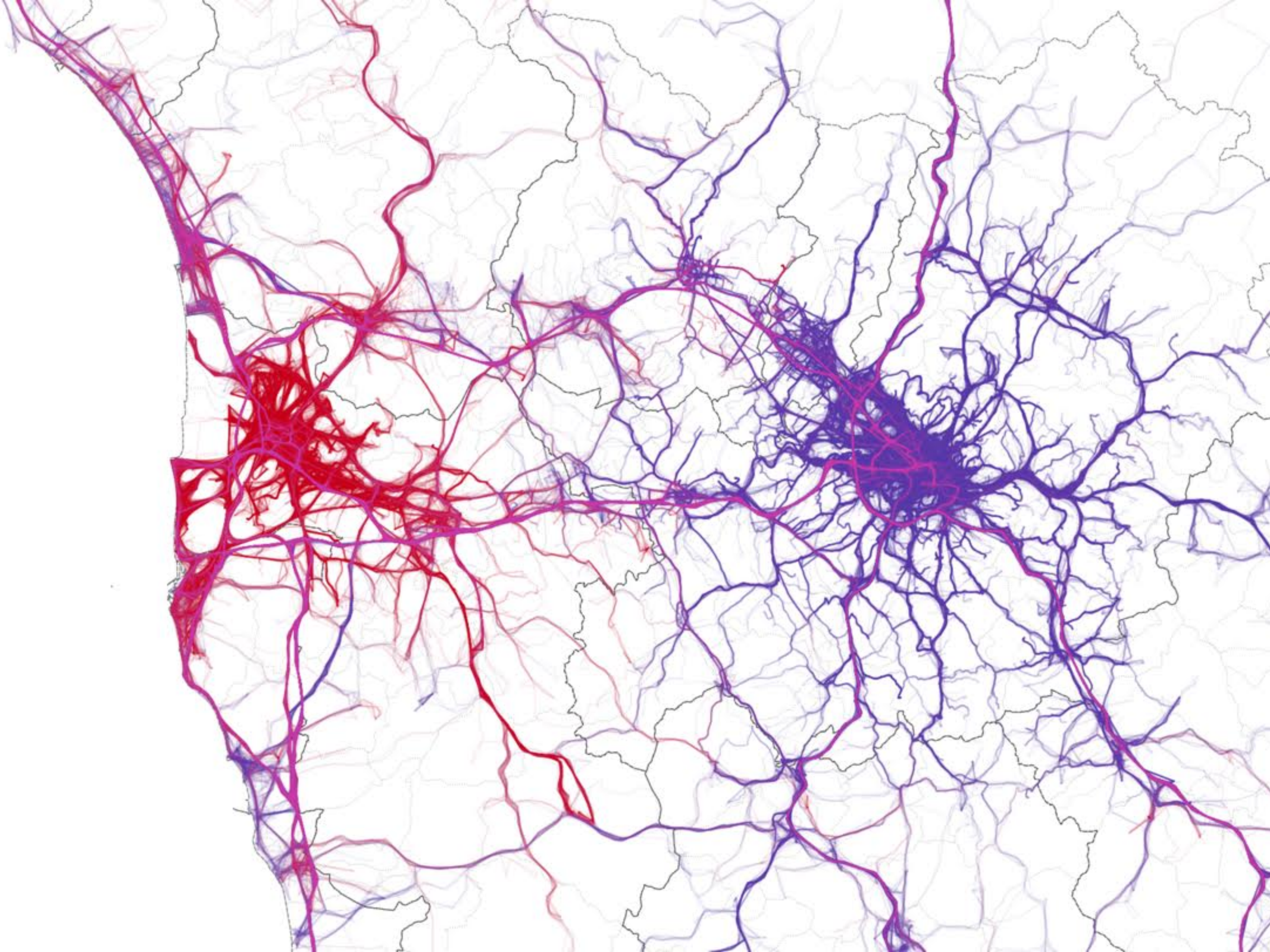
Dino Pedreschi

KDD Lab

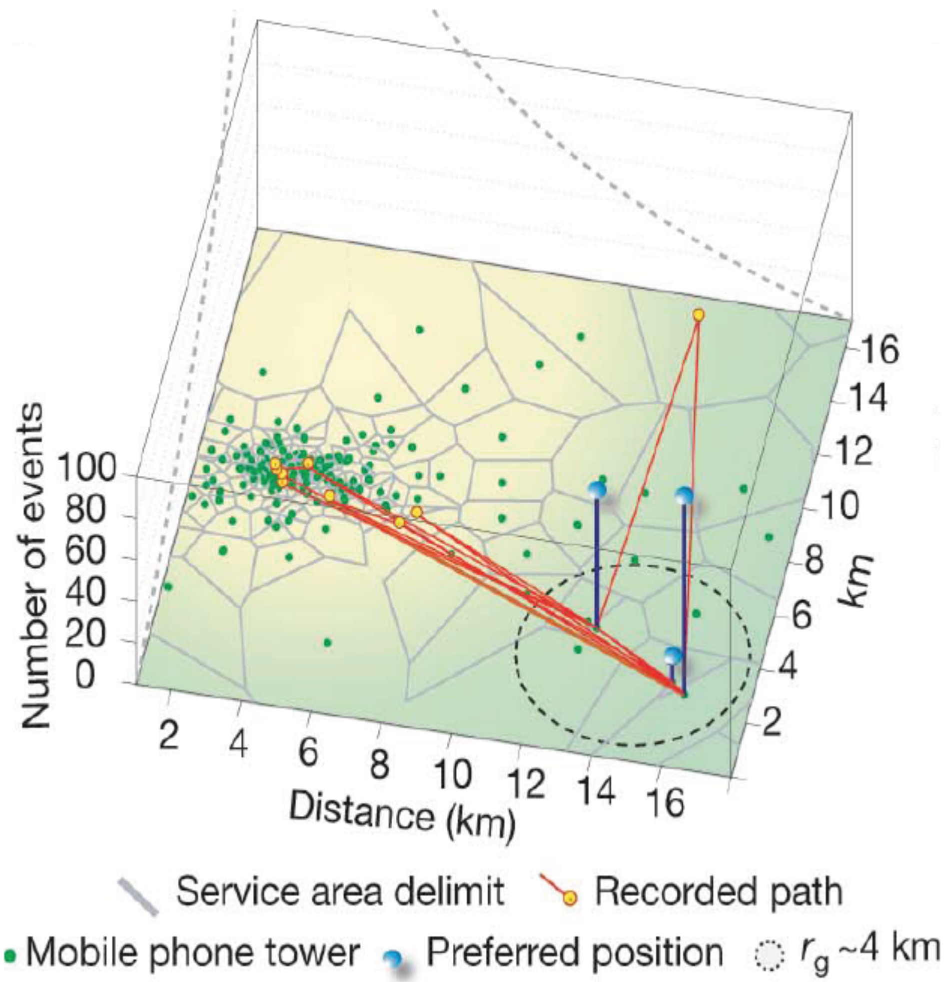
ISTITUTO DI SCIENZA E TECNOLOGIE DELL'INFORMAZIONE "A. FAEDO"

UNIVERSITÀ DI PISA

1. Pattern Discovery



mobile phone (CDR) data



when you call

where you call

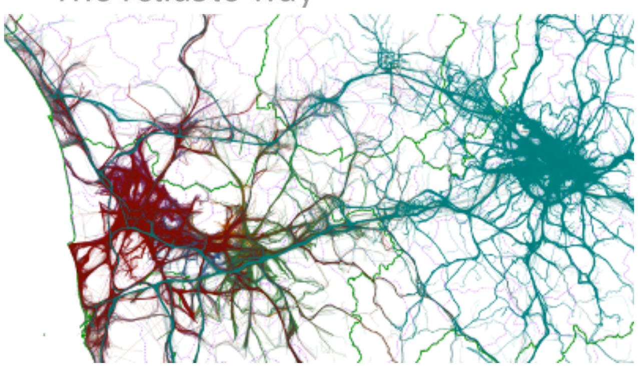
who you call

GPS and GSM data

GPS

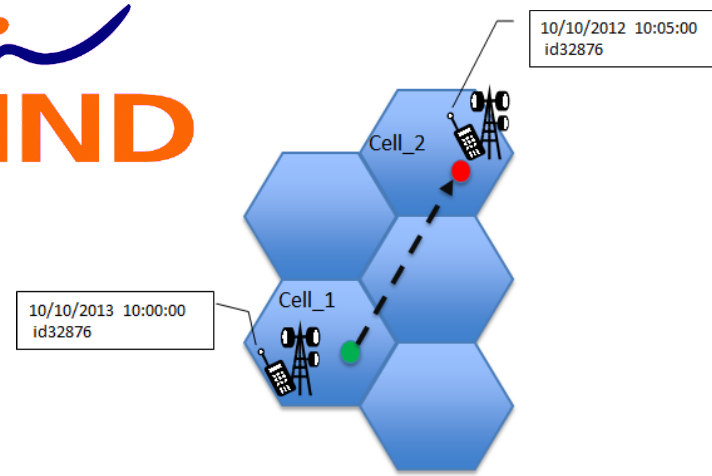
- 1 month in Tuscany
- ≈10 million car travels
- ≈ 200,000 vehicles

The reliable way



GSM

- 1 month in Tuscany
- ≈ 100 million calls
- 1 million users



nature COMMUNICATIONS

Home | About the journal | Authors and referees | Browse archive | Search

nature.com ▸ journal home ▸ current month ▸ full text

NATURE COMMUNICATIONS | ARTICLE OPEN

Returners and explorers dichotomy in human mobility

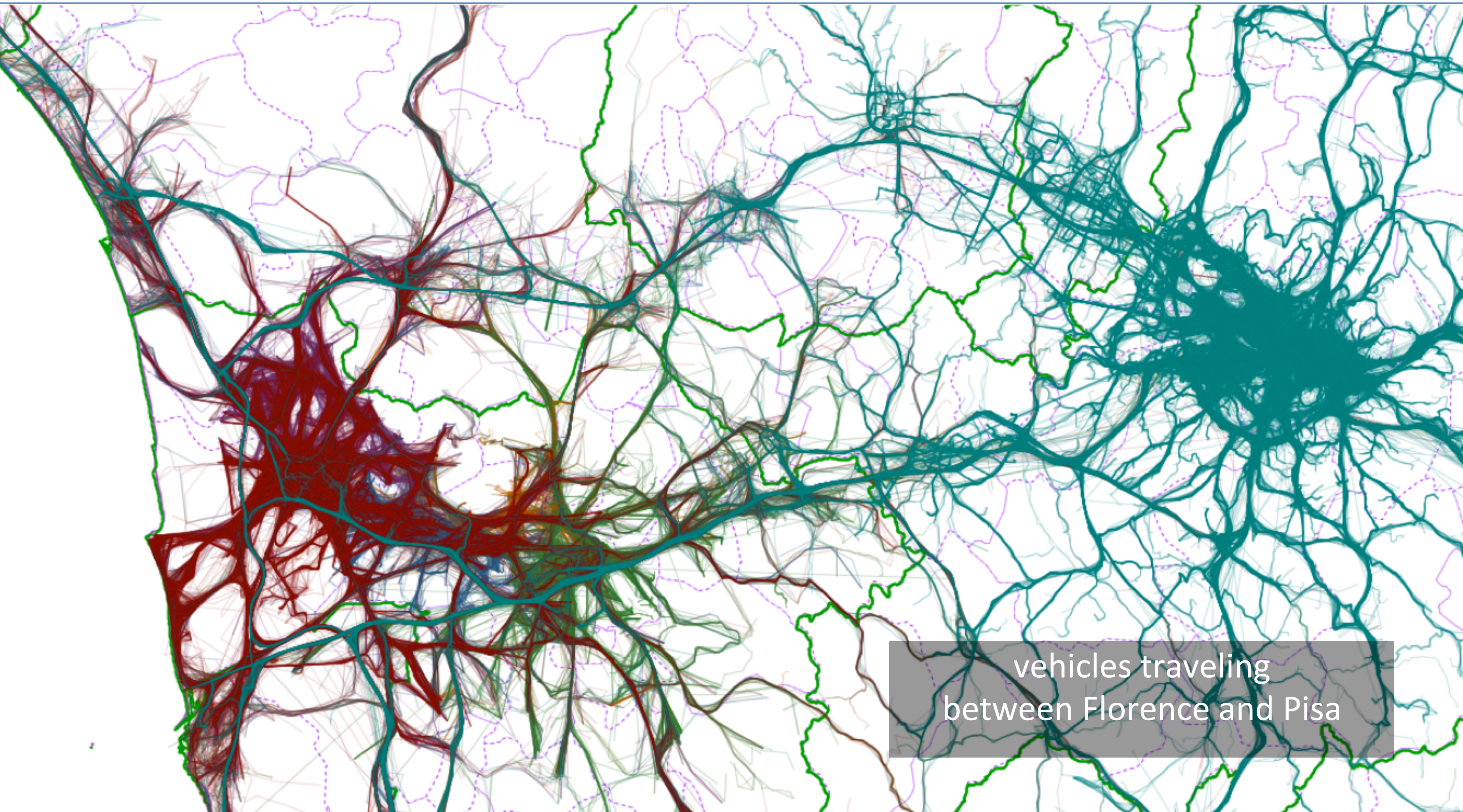
Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti & Albert-László Barabási

Affiliations | Contributions | Corresponding authors

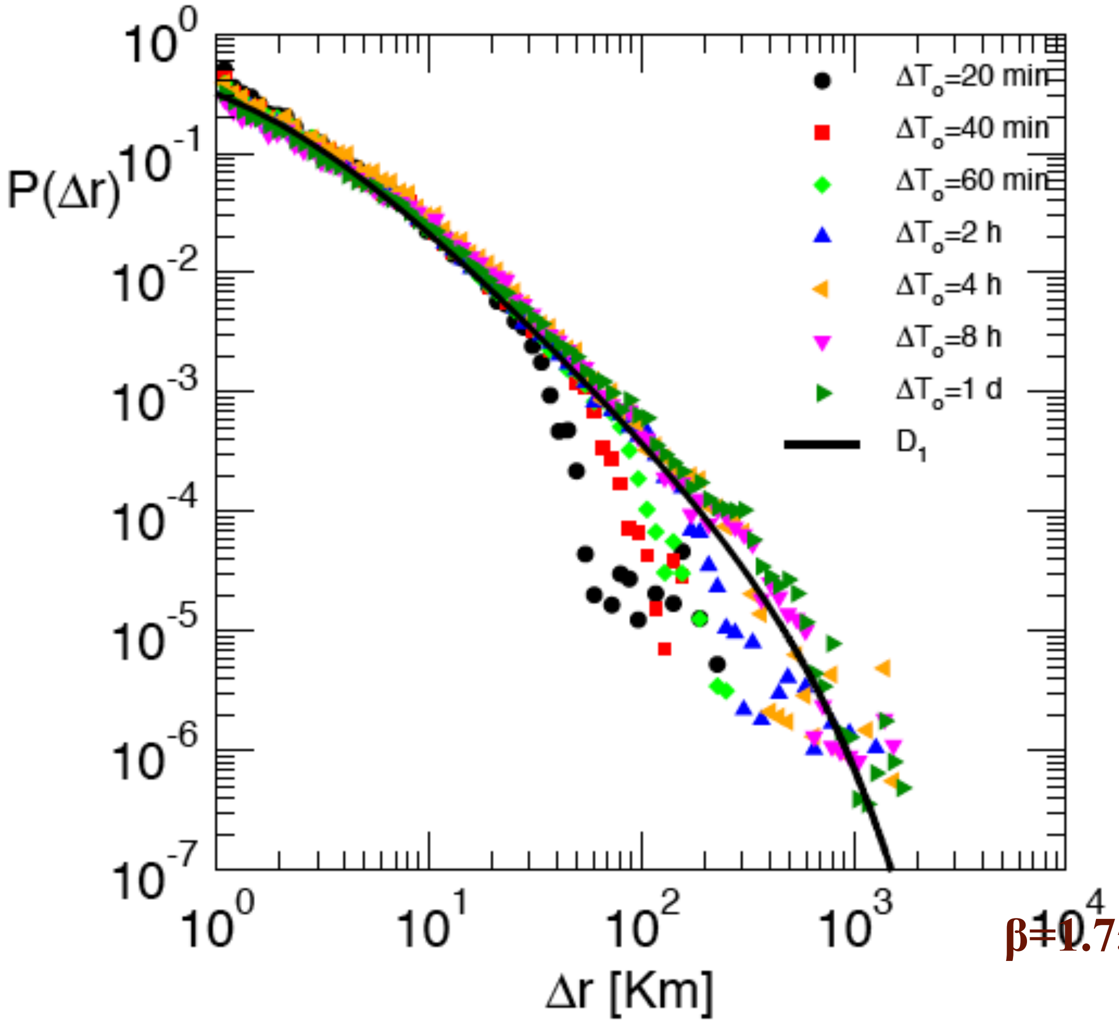
Nature Communications **6**, Article number: 8166 | doi:10.1038/ncomms9166

Received 15 December 2014 | Accepted 24 July 2015 | Published 08 September 2015

Human mobility is a complex system



Scale-free distribution of travel length

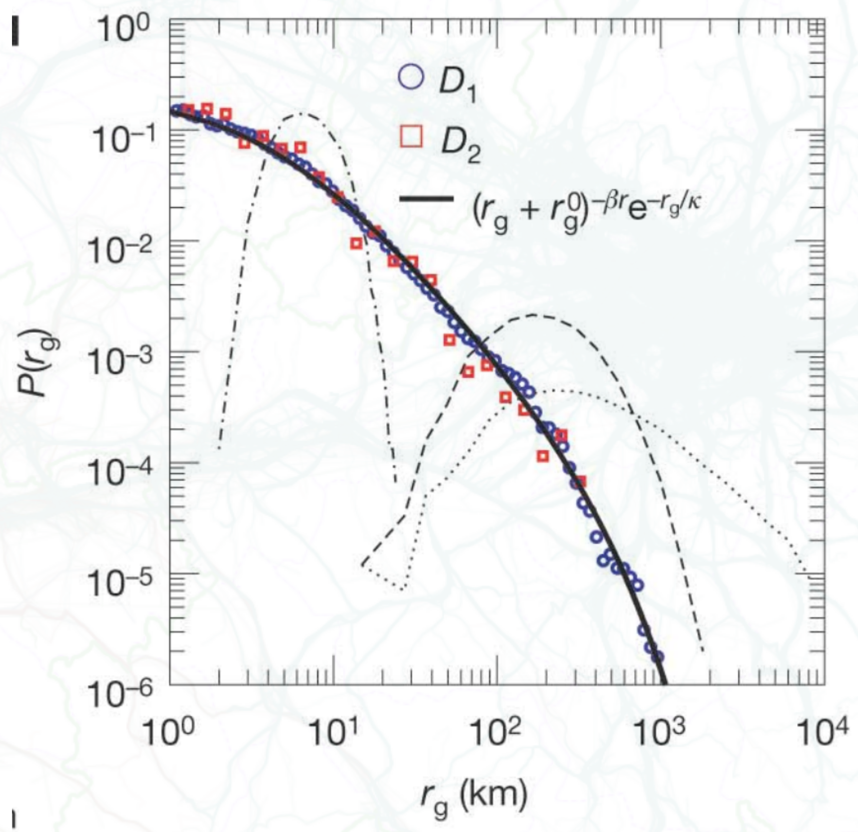


β=1.75±0.15

Candia, González *et al.*
J. Phys. A: Math. Theor. **41** (2008)

The heterogeneity of human mobility

$$r_g = \sqrt{\frac{1}{N}\sum_{i \in L} n_i(\vec{r}_i - \vec{r}_{cm})^2},$$

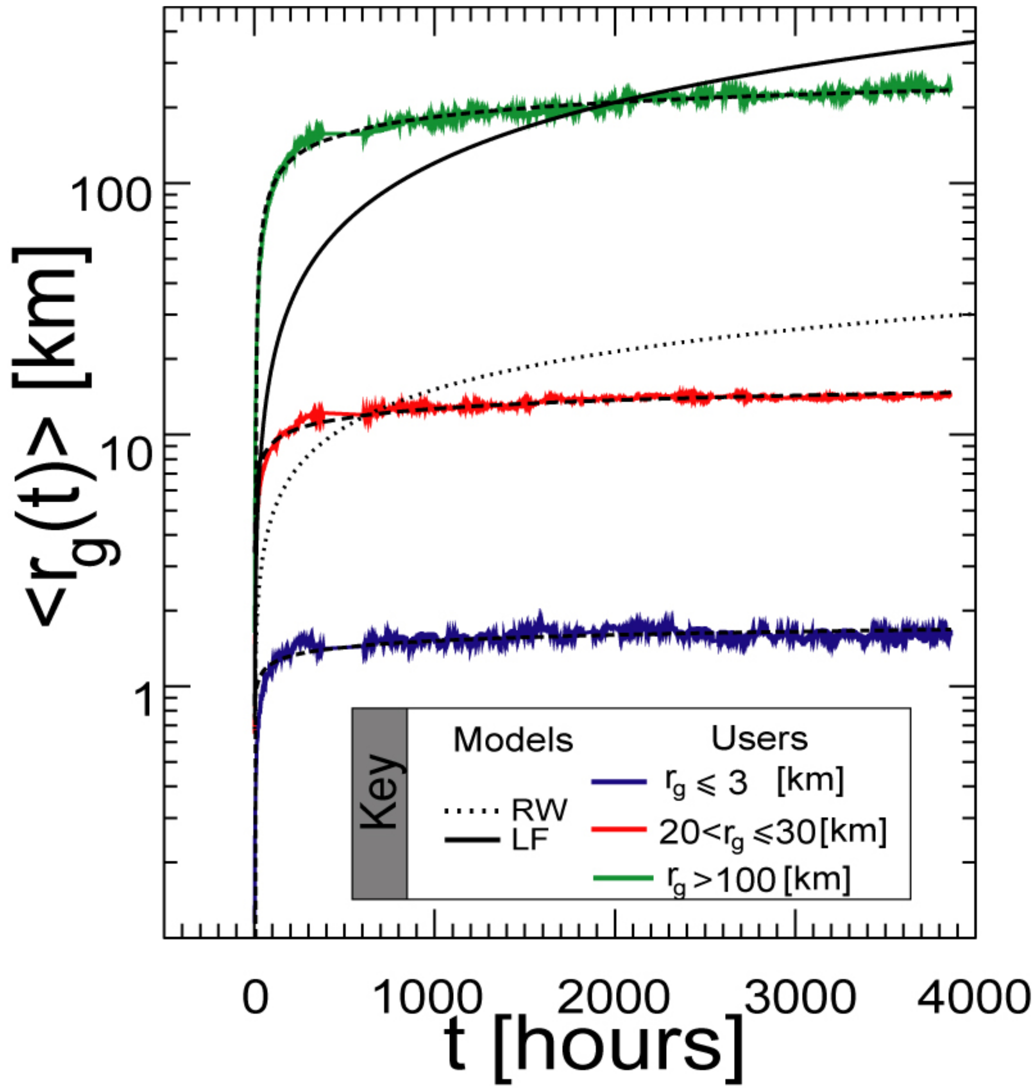


Characteristic traveled distance

Radius of Gyration:

$$r_g^a(t) = \sqrt{\frac{1}{n_c^a(t)} \sum_{i=1}^{n_c^a} (\vec{r_i^a} - \vec{r}_{cm}^a)^2}$$

$$\vec{r}_{cm} = \frac{1}{n_p} \sum_{i=1}^{n_p} \vec{r_i}.$$

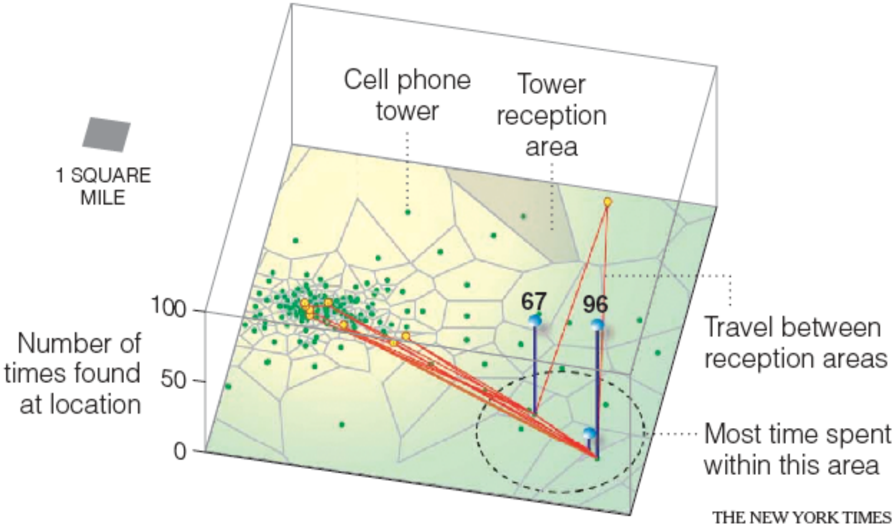


Recurrent mobility

What is the impact of *recurrent* mobility on total mobility of individuals?

mobile phone data

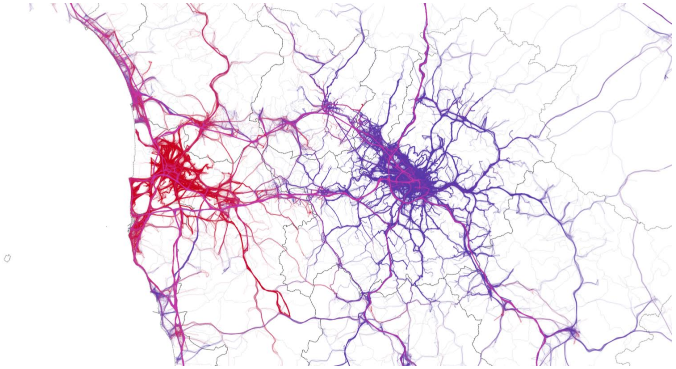
- 67,000 users
- a big country
- 3 months



GPS tracks

- 40,000 vehicles
- Tuscany
- 1 month

locations = census cells



total vs recurrent mobility

- **total radius**

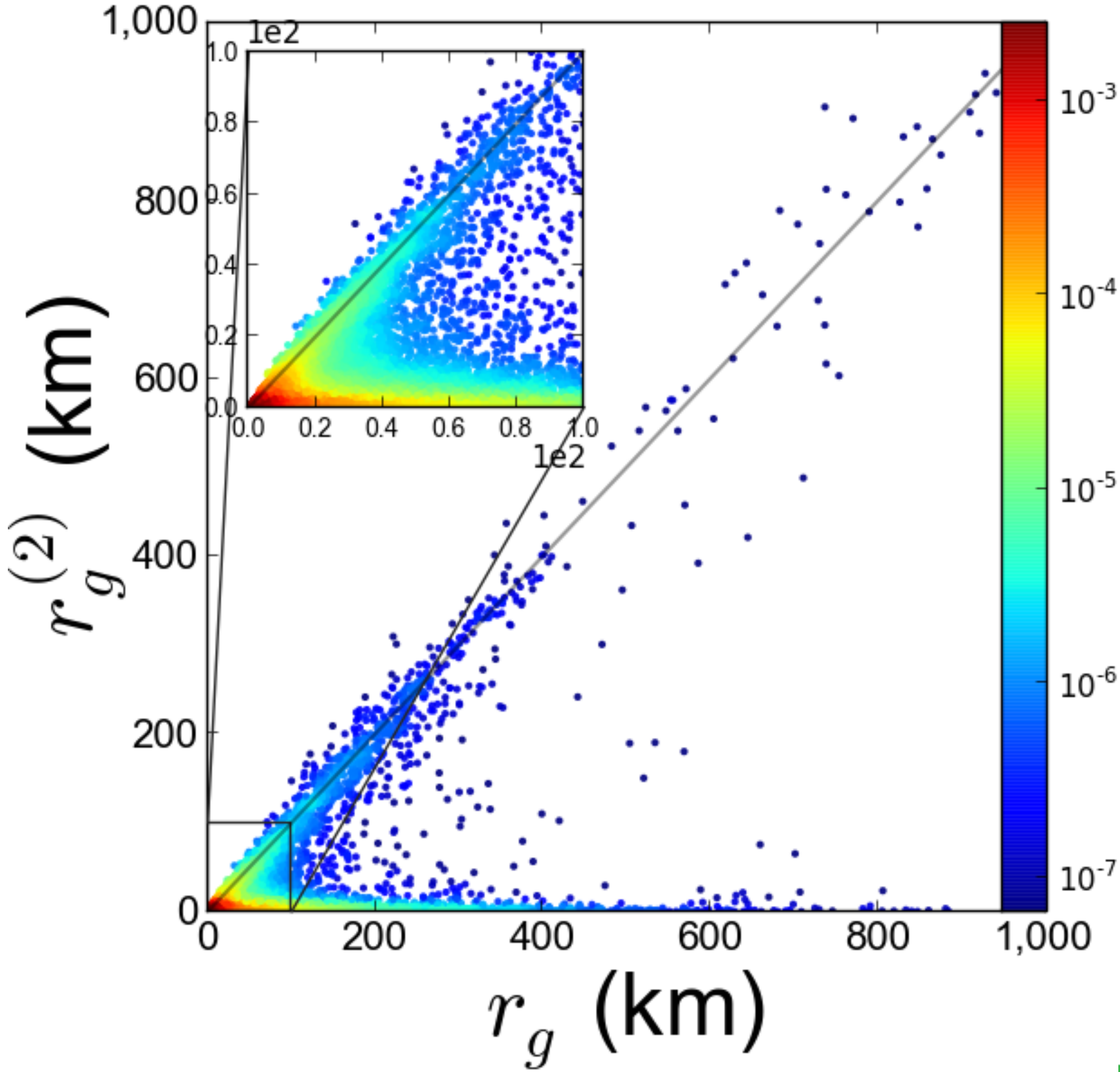
$$r_g = \sqrt{\frac{1}{N} \sum_{i \in L} n_i (\vec{r}_i - \vec{r}_{cm})^2},$$

the characteristic distance traveled by individuals

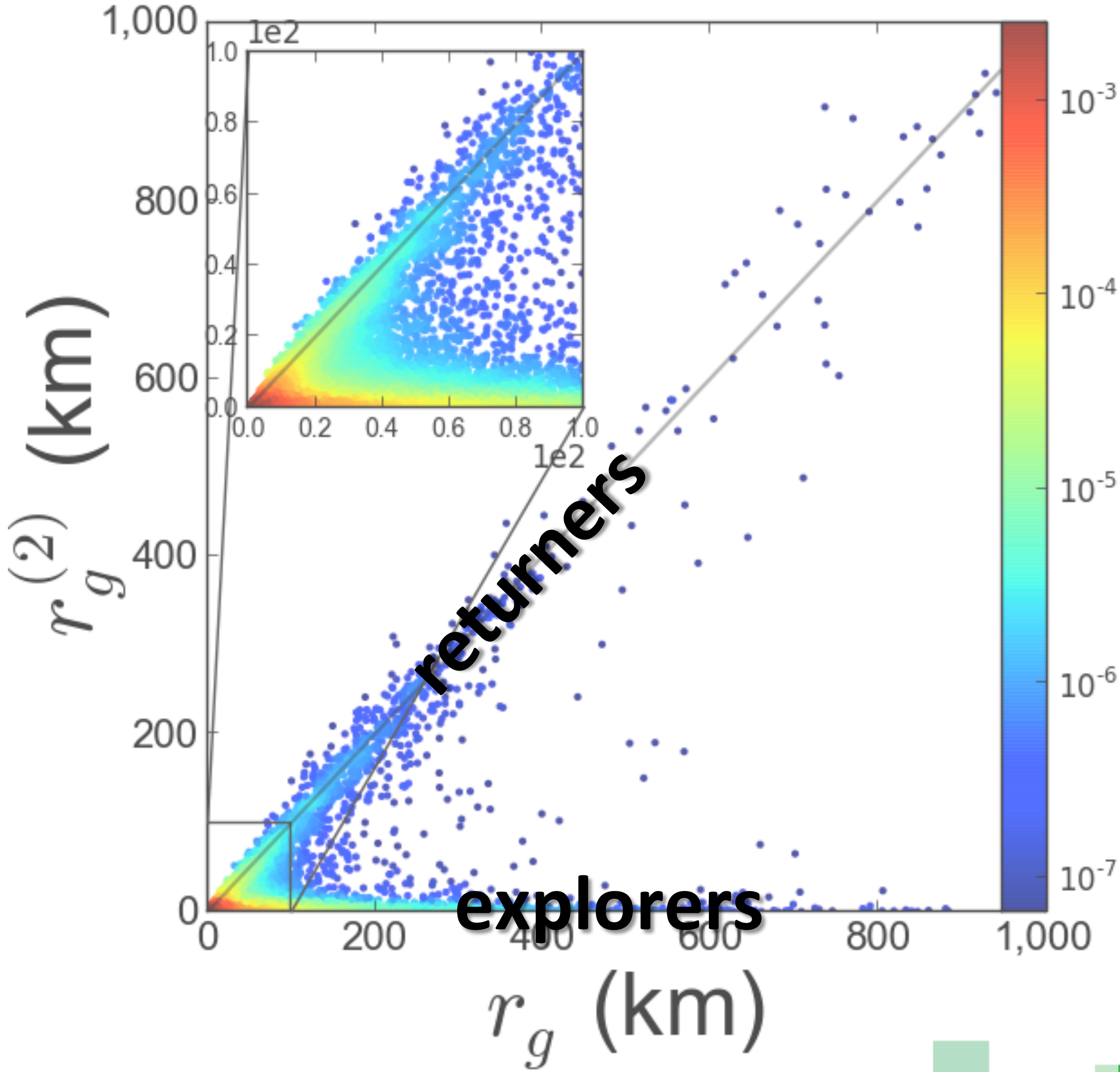
recurrent radius

$$r_g^{(k)} = \sqrt{\frac{1}{N_k} \sum_{i=1}^{k} n_i (\vec{r}_i - \vec{r}_{cm}^{(k)})^2}$$

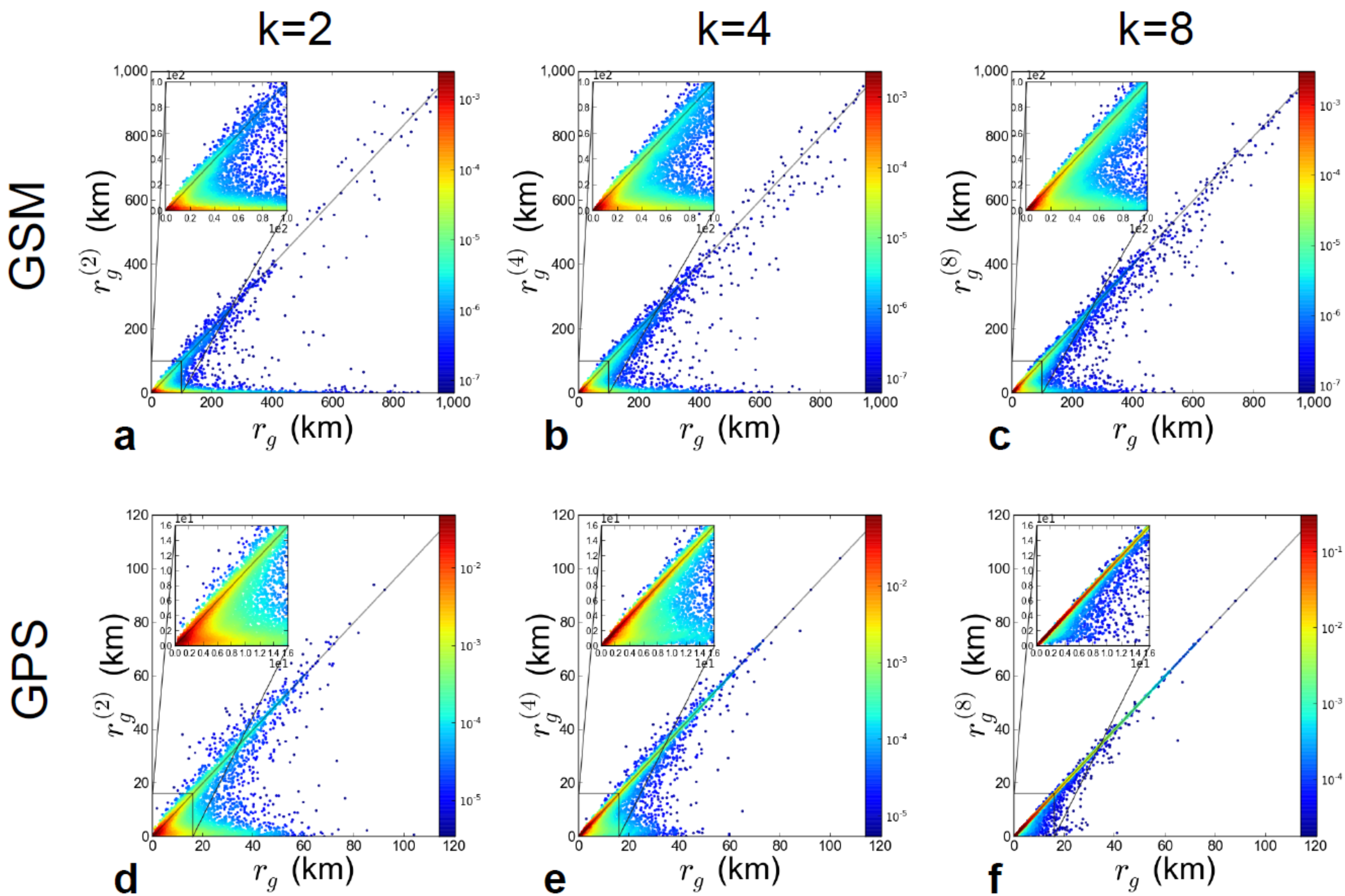
the radius computed on the k most visited locations

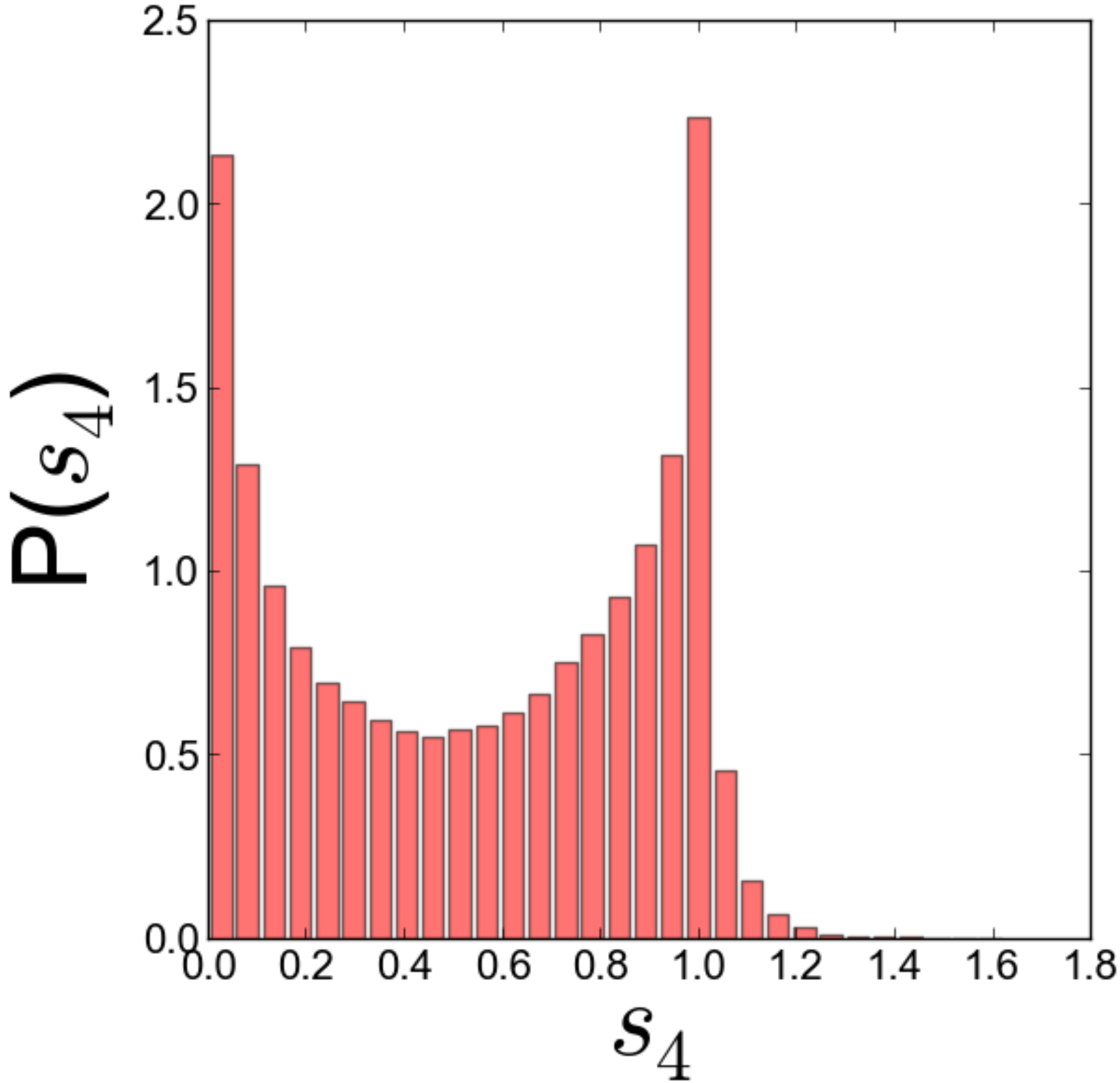


1,000
800
600
400
200
0
0
200
400
600
800
1,000
$r_g^{(2)}$ (km)
r_g (km)
1e2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1e2
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}

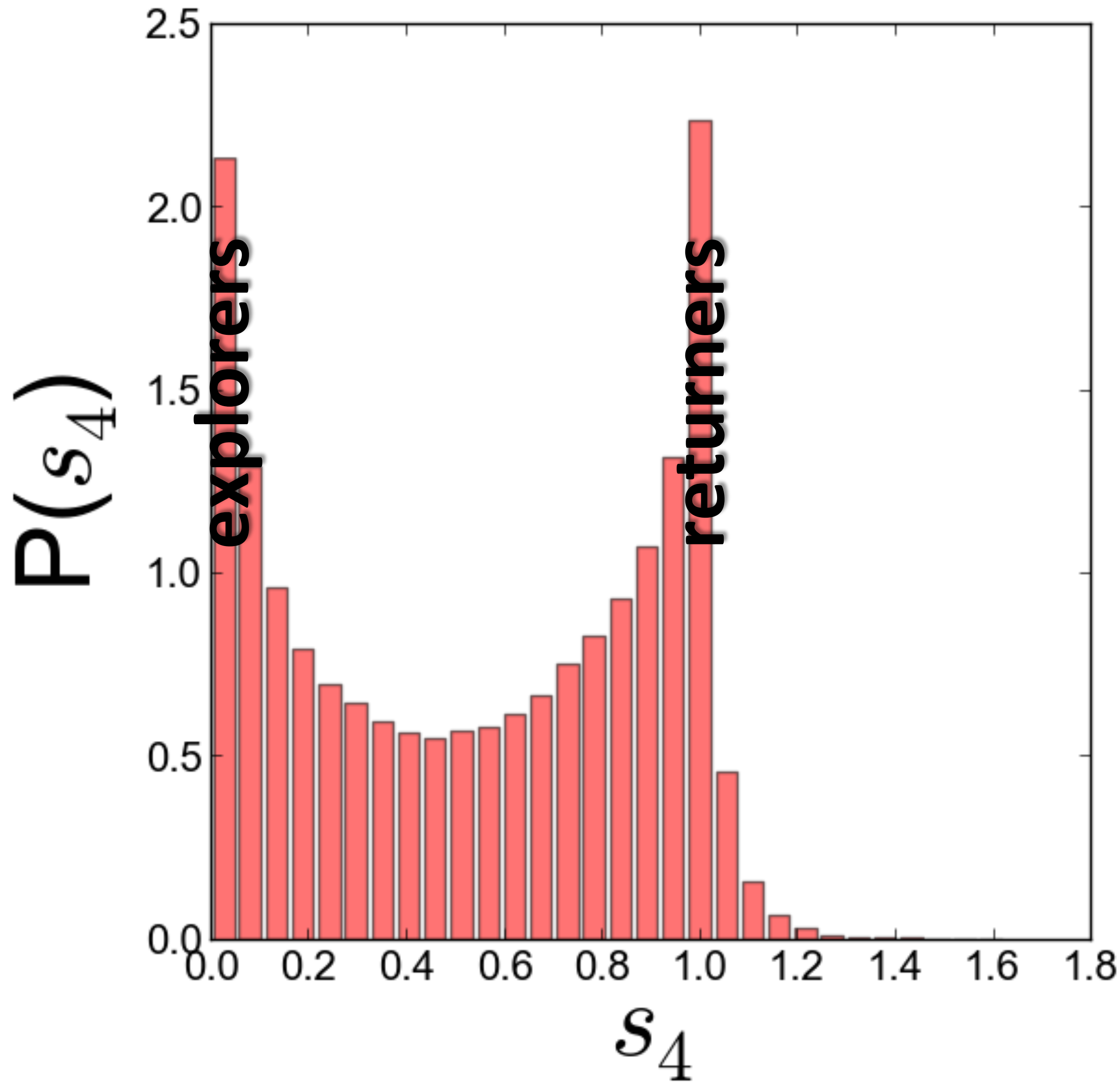


returners
explorers
$r_g^{(2)}$ (km)
r_g (km)
1e2

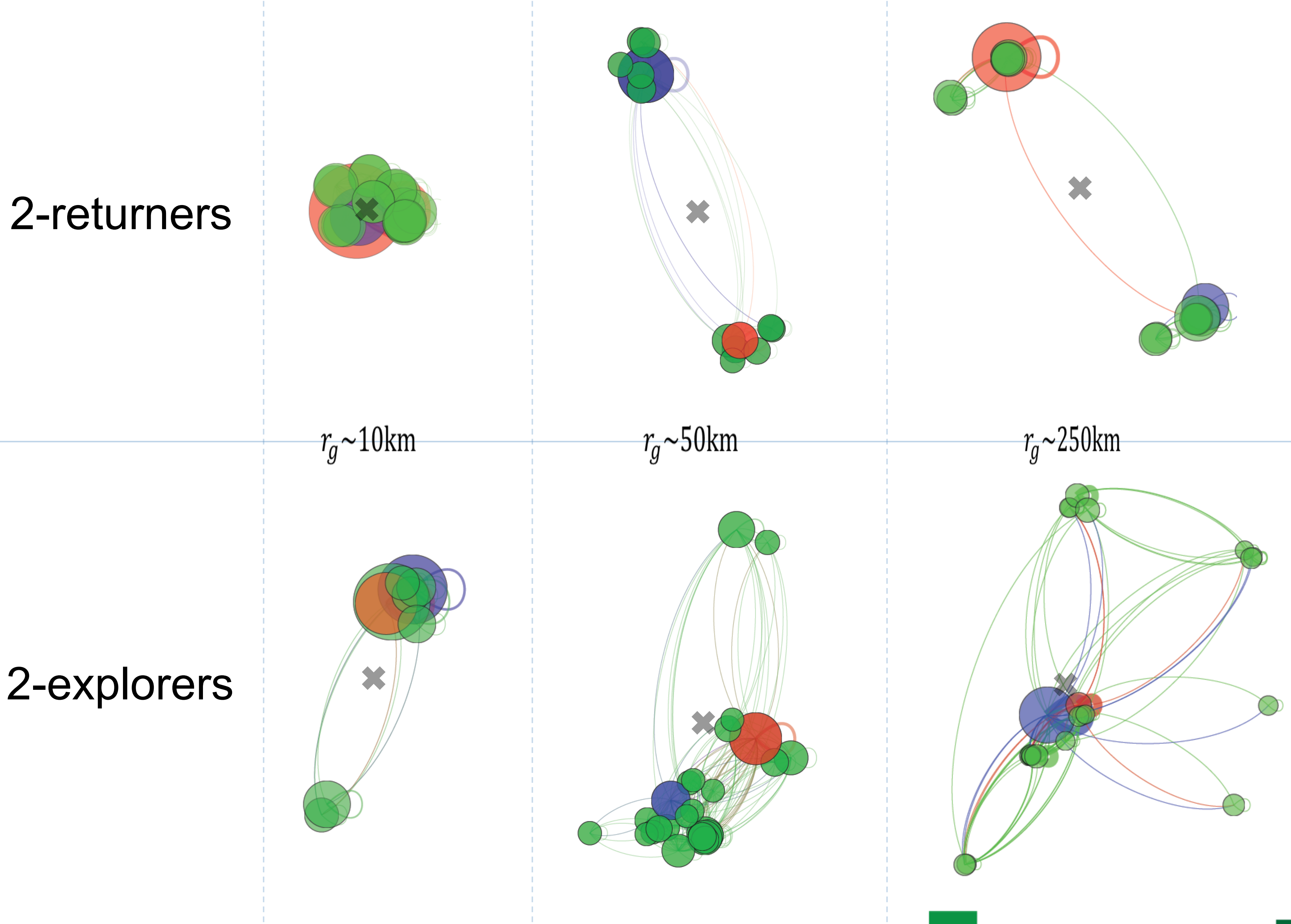

k=2
k=4
k=8
GSM
GPS
$r_g^{(2)}$ (km)
$r_g^{(4)}$ (km)
$r_g^{(8)}$ (km)
r_g (km)
a
b
c
d
e
f



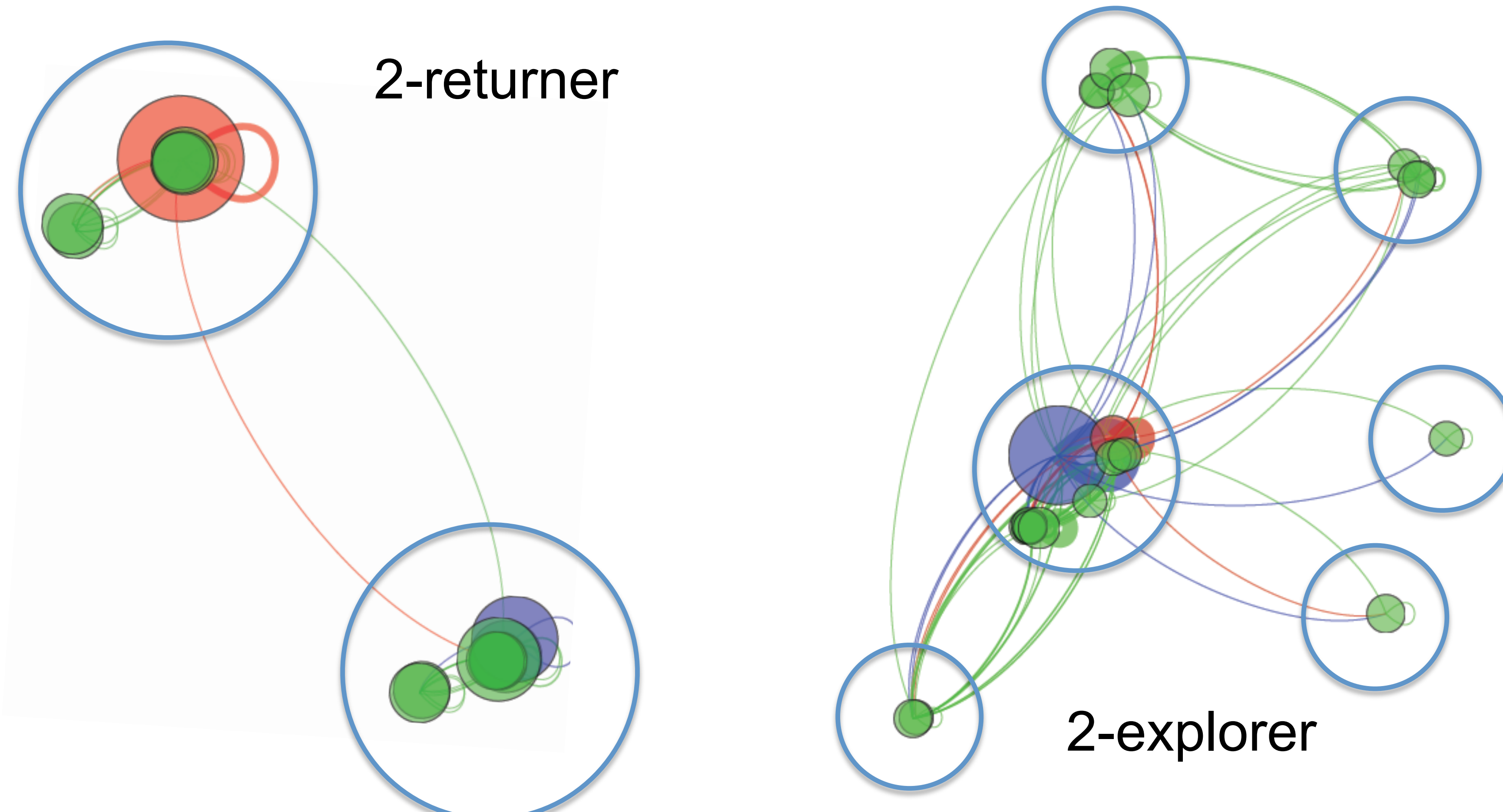
P(s_4)
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
s_4



explorers
returners
$P(s_4)$
s_4
2.5
2.0
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8


2-returners
r_g~10km
r_g~50km
r_g~250km
2-explorers

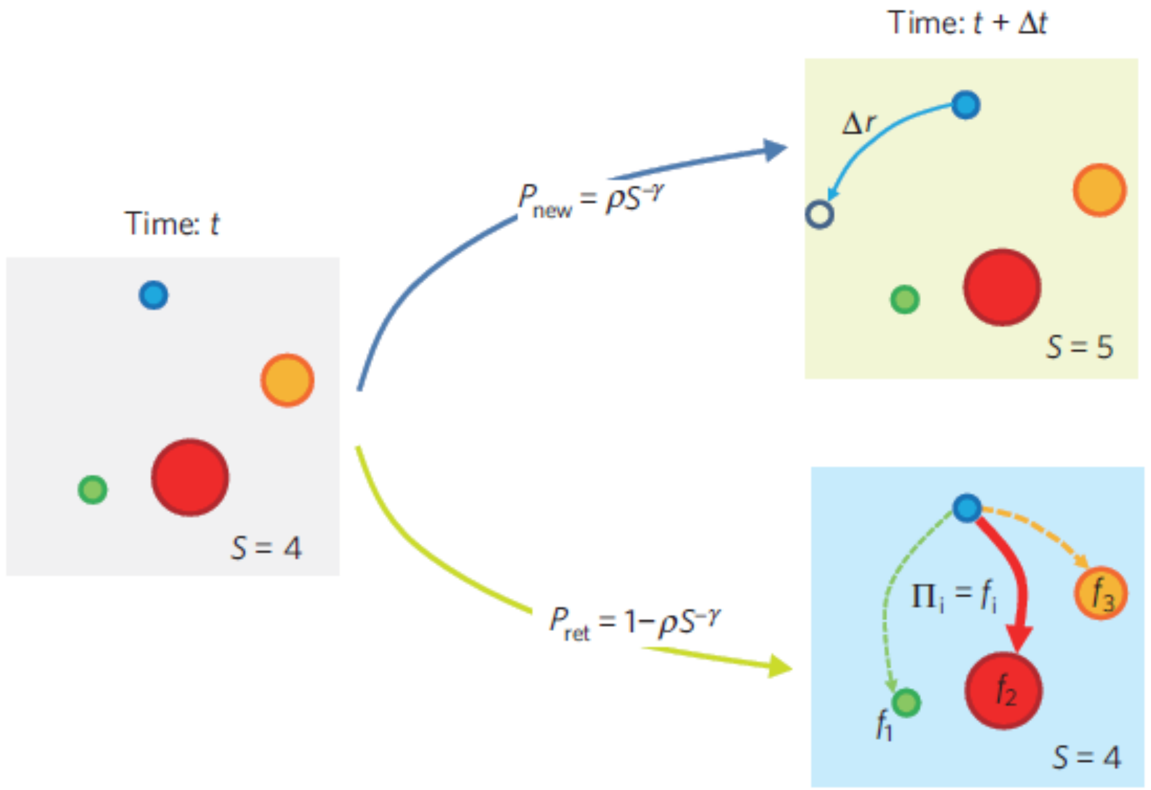
Spatial clusters



1) **Clusters law**:
in individual mobility networks locations tend to aggregate in dense clusters

2) **Returners/Explorers law**:
as total rg increases the k most frequent locations move far apart for returners, they remain close for explorers.

EPR model of human mobility

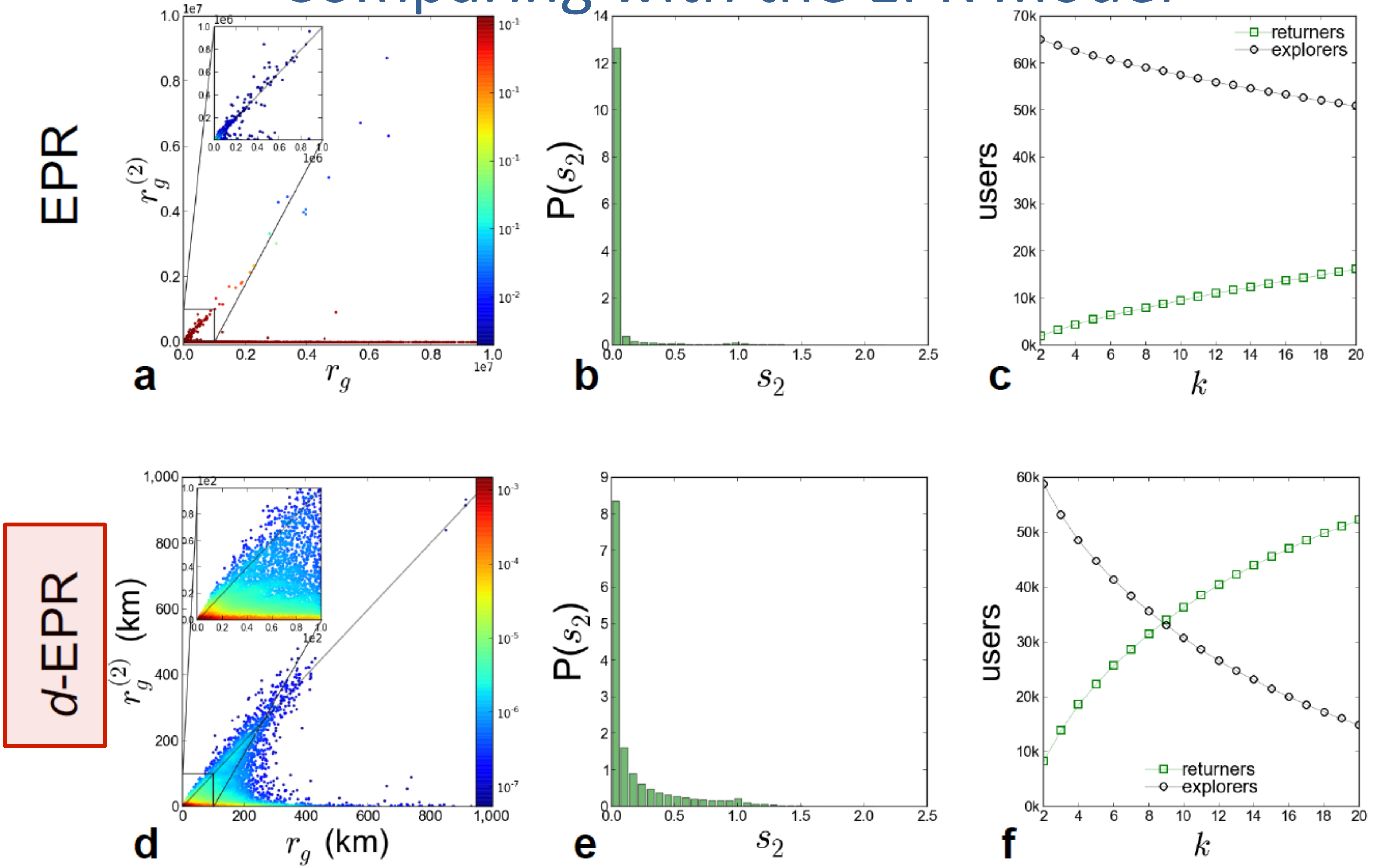


- **Exploration**
- **Preferential return**

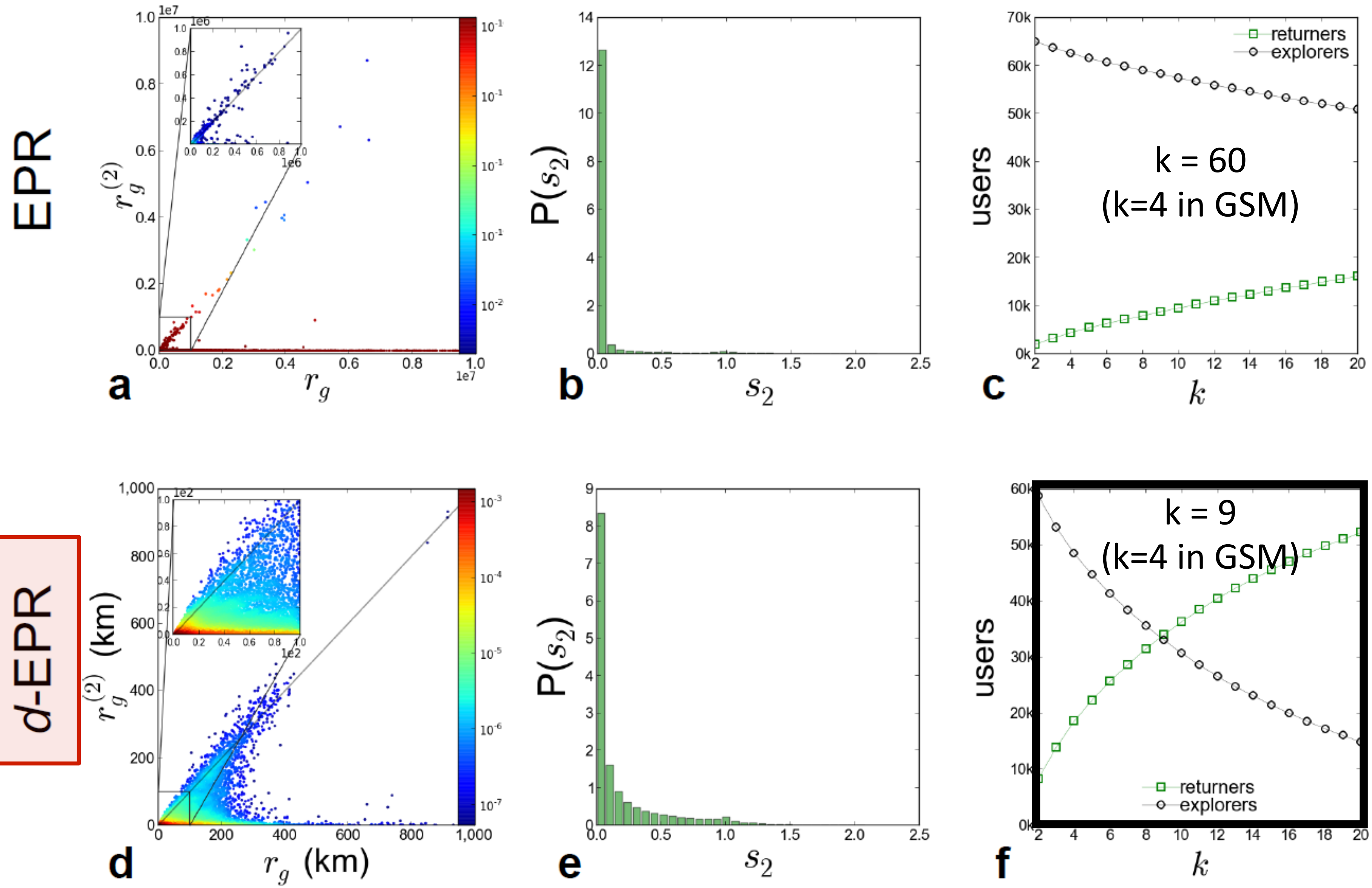
A theoretical basis for agent-based simulation of population dynamics

Song, Koren, Wang, Barabasi, *Nature Physics, Sept.* 2010

Comparing with the EPR model

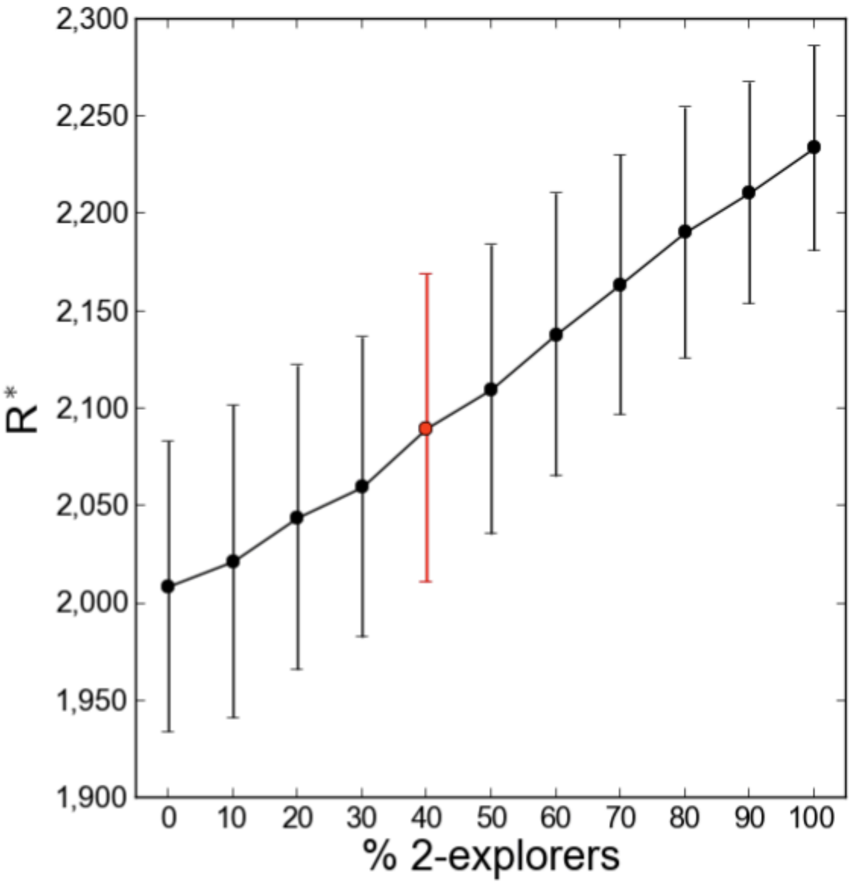


COMPARING WITH THE EPR MODEL



Exploreres and Diffusion

- **The global invasion diffusion threshold**.
- The bars show how the distribution of the *diffusion invasion threshold* changes when different proportions of returners and explorers are chosen.
- The red bar indicates the distribution where the fraction of explorers is 40%, the actual fraction of explorers in real data.

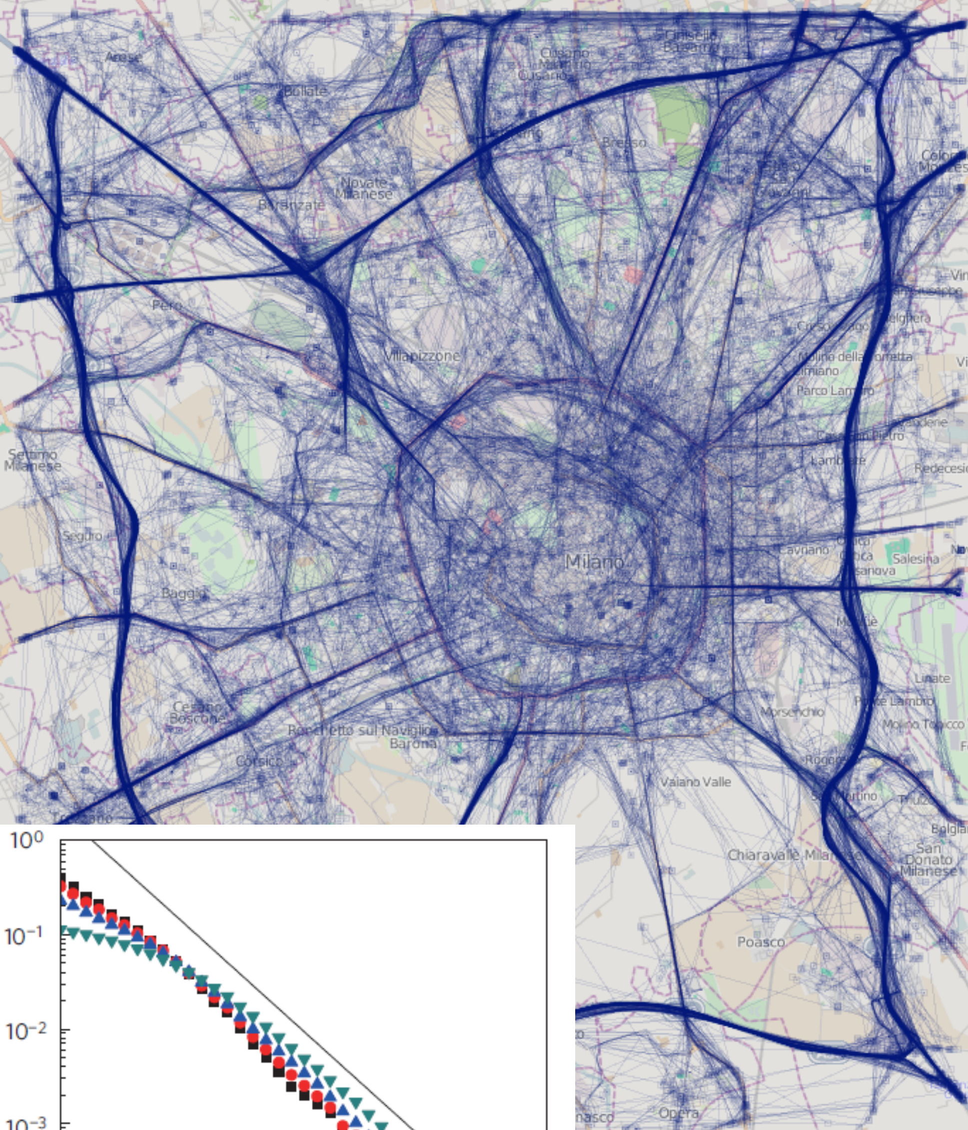


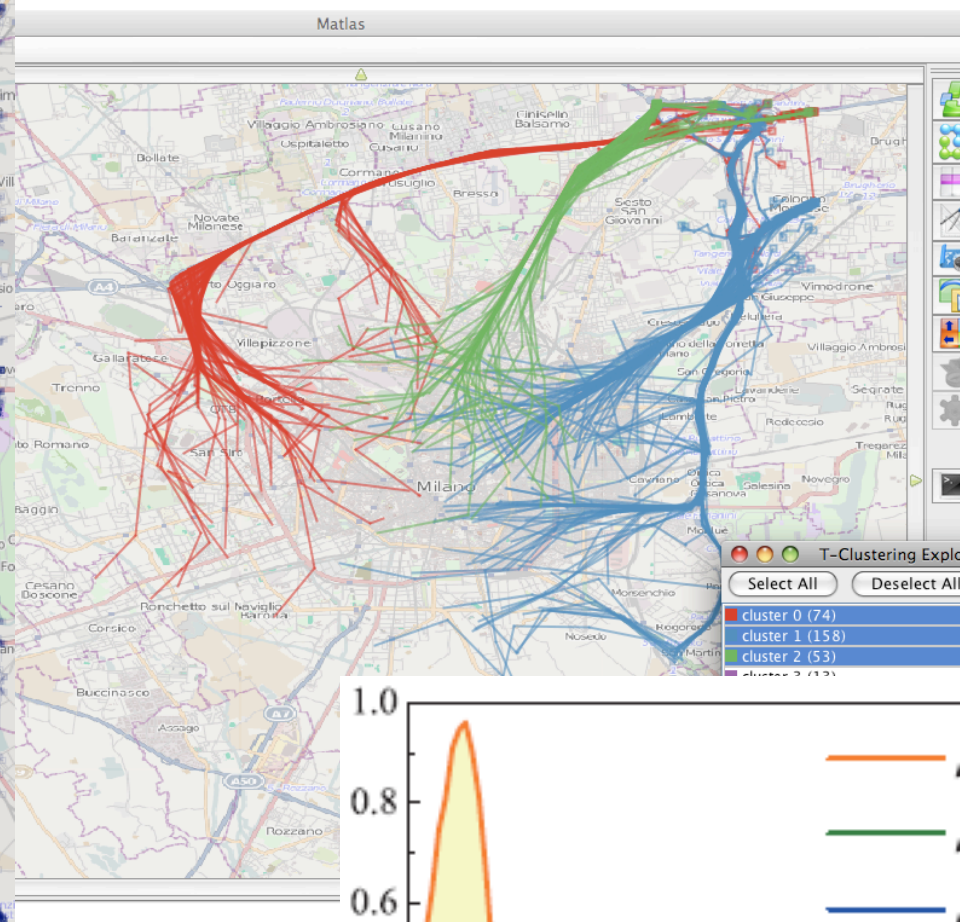
FINDINGS

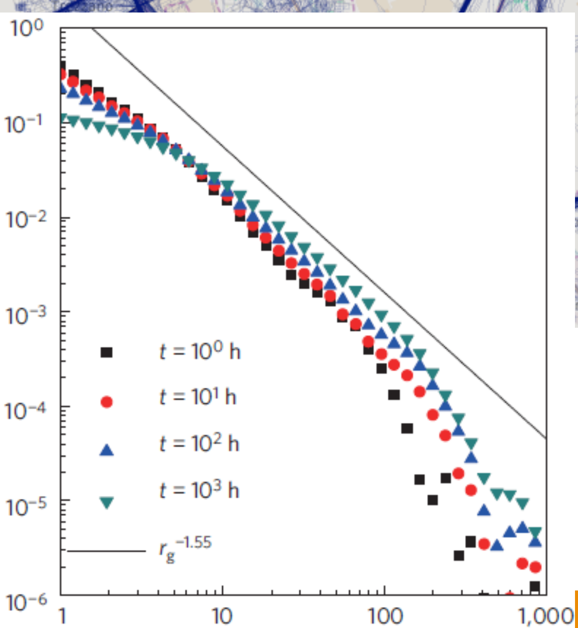
1. Returners and explorers are sharply separate profiles
2. Explorers are crucial actors in the diffusion of diseases or other spreading phenomena
3. Social Homophily: returners preferably call other returners (the same applies to explorers)

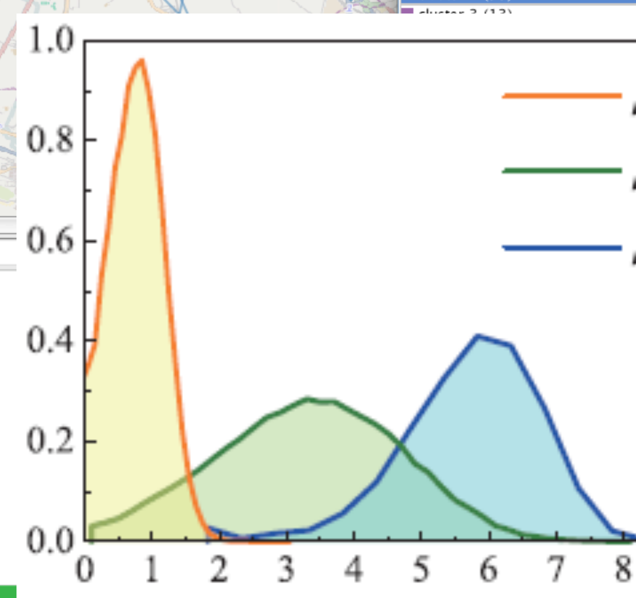
Big data push towards convergence

- Network science / complex system science
  - **Global models** of complex social phenomena
  - Behavioral **diversity** in society at large
- Data mining
  - **Local patterns** of complex social phenomena
  - Behavioral **similarity** in sub-populations
- Convergence needed to achieve realistic and accurate models for prediction and **simulation**


Milano
Bollate
Novate Milanese
Villapizzone
Baggio
Poasco
Opera


Matlas
Select All
cluster 0 (74)
cluster 1 (158)
cluster 2 (53)
Milano
Bollate
Villapizzone
Trenno
Rozzano
Buccinasco

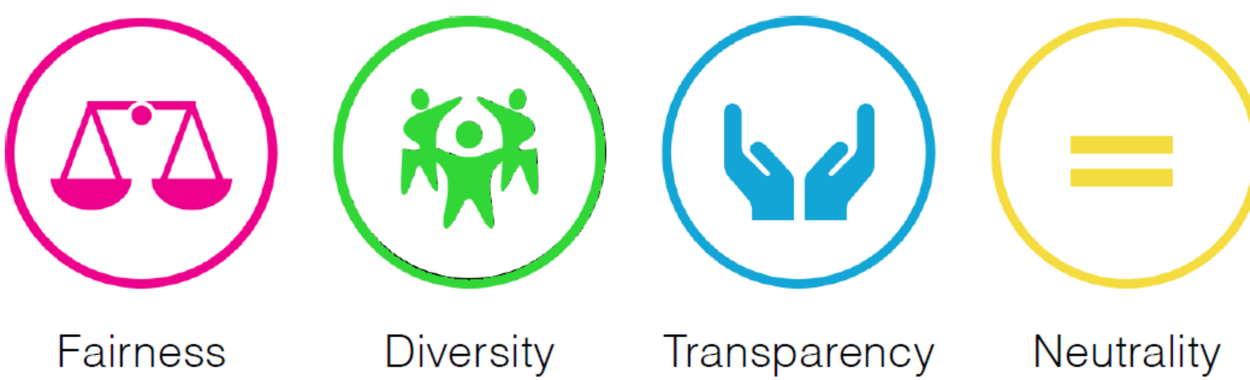

10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
$t = 10^0$ h
$t = 10^1$ h
$t = 10^2$ h
$t = 10^3$ h
$r_g^{-1.55}$
1
10
100
1,000


1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
7
8

2. Machine Learning Transparency

Big Data, Big Risks

- **Big data is algorithmic, therefore it cannot be biased!** And yet…
- All traditional evils of social discrimination, and many new ones, exhibit themselves in the big data ecosystem
- Because of its tremendous **power**, massive data analysis must be used **responsibly**
- Technology alone won't do: also need **policy**, **user involvement** and **education** efforts



- By 2018, 50% of business ethics violations will occur through improper use of big data analytics

- [source: Gartner, 2016]

The danger of black boxes

- The COMPAS score (Correctional Offender Management Profiling for Alternative Sanctions)
- A 137-questions questionnaire and a predictive model for "risk of crime recidivism." The model is a proprietary secret of Northpointe, Inc.

- The data journalists at propublica.org have shown that the model has a strong ethnic bias
  - blacks who did not reoffend are classified as high risk twice as much as whites who did not reoffend
  - whites who did reoffend were classified as low risk twice as much as blacks who did reoffend.

The danger of black boxes

- The three major US credit bureaus, Experian, TransUnion, and Equifax, providing credit scoring for millions of individuals, are often discordant.

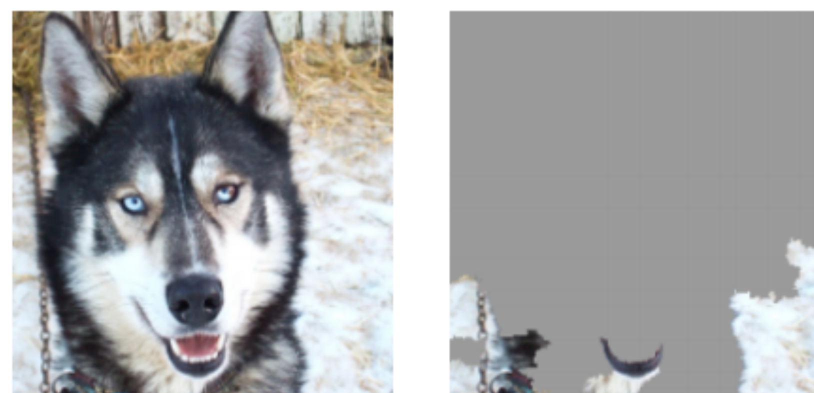
- In a study of 500,000 records, 29% of consumers received credit scores that differ by at least fifty points between credit bureaus, a difference that may mean tens of thousands dollars over the life of a mortgage [CRS+16].

The danger of black boxes

- An accurate but untrustworthy classifier may result from an accidental bias in the training data.
- In a task of discriminating wolves from huskies in a dataset of images, the resulting deep learning model is shown to classify a wolf in a picture based solely on …

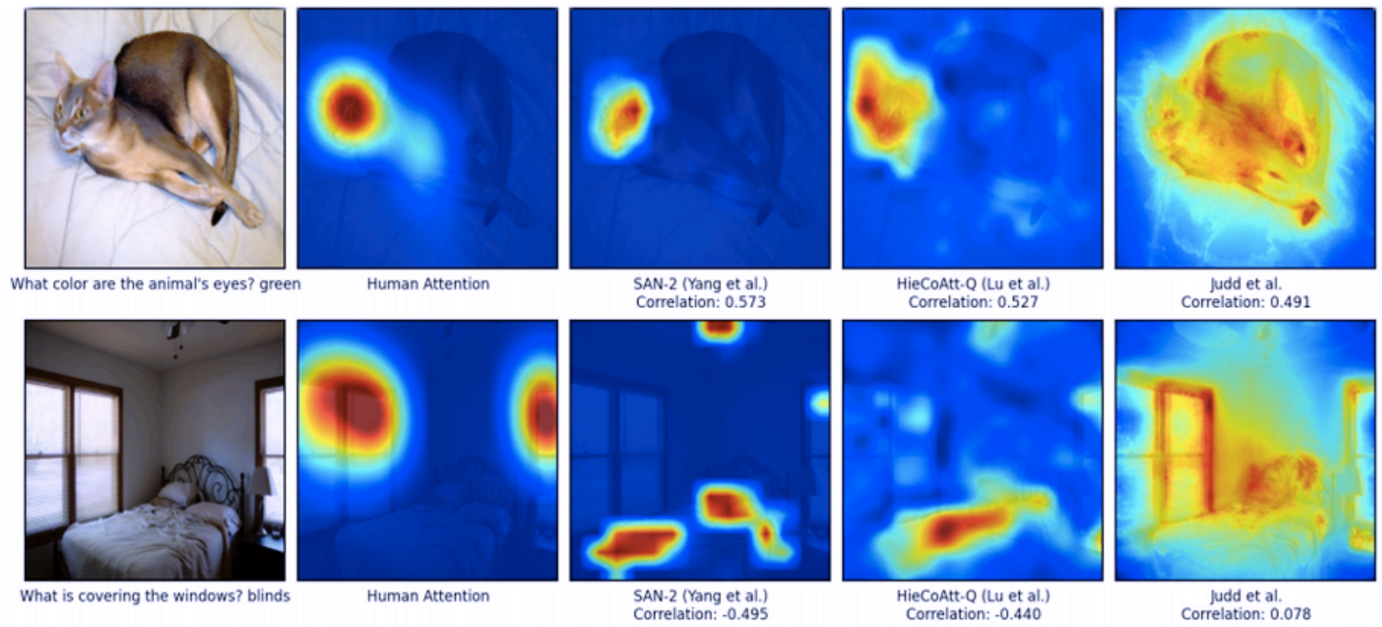
The danger of black boxes

- An accurate but untrustworthy classifier may result from an accidental bias in the training data.
- In a task of discriminating wolves from huskies in a dataset of images, the resulting deep learning model is shown to classify a wolf in a picture based solely on … **the presence of snow in the background!**



(a) Husky classified as wolf (b) Explanation

Deep learning is creating computer systems we don't fully understand



"THEY'RE PICKING [ANSWERS] BASED ON BIASES IN THE DATA SETS, RATHER THAN FROM FACTS ABOUT THE WORLD."

- As we stated in our 2008 SIGKDD paper that started the field of discrimination-aware data mining [PRT08]:

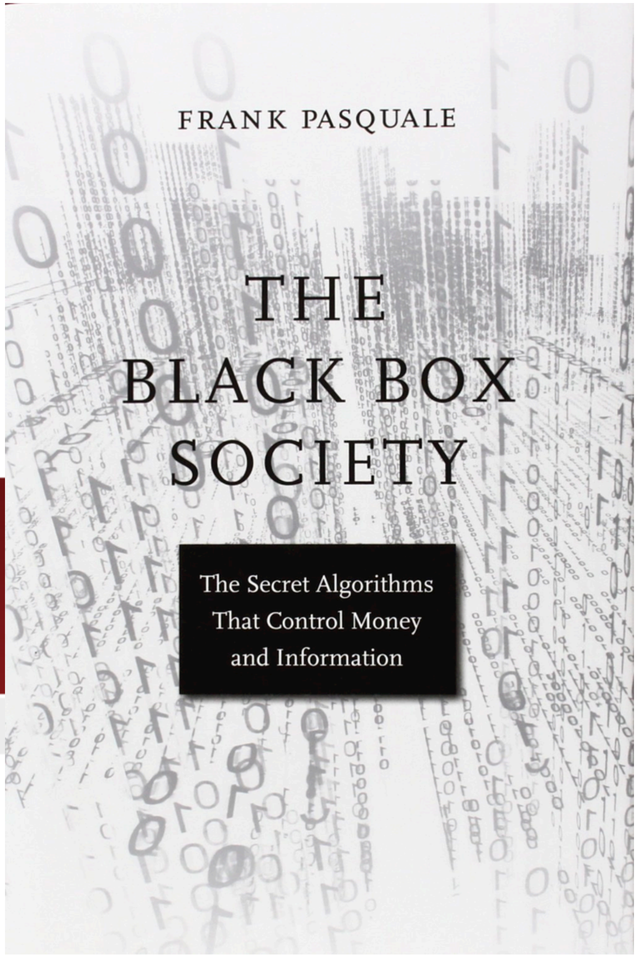
- “learning from historical data recording human decision making may mean to discover traditional prejudices that are endemic in reality, and to assign to such practices the status of general rules, maybe unconsciously, as these rules can be deeply hidden within the **Discrimination-aware Data Mining**

Dino Pedreschi Salvatore Ruggieri Franco Turini
Dipartimento di Informatica, Università di Pisa
L.go B. Pontecorvo 3, 56127 Pisa, Italy
{pedre,ruggieri,turini}@di.unipi.it

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

Transparent algorithms to build trust

- **Systems that recommend humans making a decision should explain why**



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

Archive > Volume 537 > Issue 7621 > Editorial > Article

NATURE | EDITORIAL

More accountability for big-data algorithms

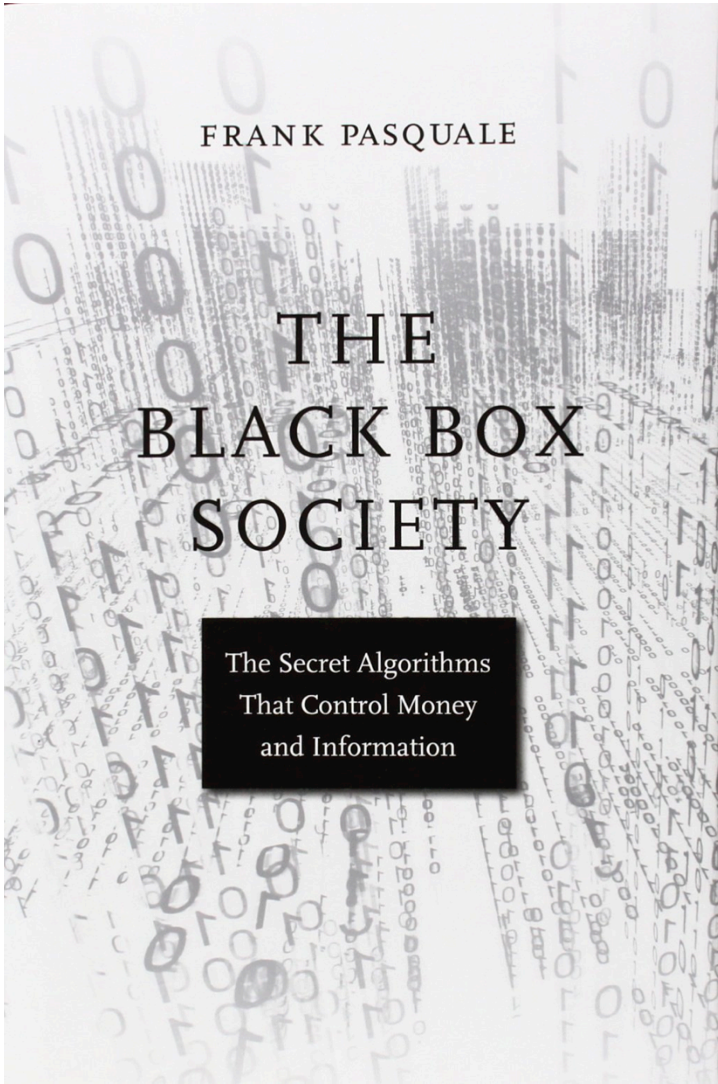
To avoid bias and improve transparency, algorithm designers must make data sources and profiles public.

21 September 2016

Right of explanation

- Applying AI within many domains requires **transparency** and **responsibility**:
    - health care
    - finance
    - surveillance
    - autonomous vehicles
    - Government

- EU General Data Protection Regulation (April 2016) establishes a right of explanation for all individuals to obtain "meaningful explanations of the logic involved" when automated (algorithmic) individual decision-making, including profiling, takes place.



Data ethics technologies

Discrimination discovery from data

Discrimination-aware Data Mining

Dino Pedreschi Salvatore Ruggieri Franco Turini
Dipartimento di Informatica, Università di Pisa
L.go B. Pontecorvo 3, 56127 Pisa, Italy
{pedre,ruggieri,turini}@di.unipi.it

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

Discrimination discovery

- Given:
  - an historical database of **decision records**, each describing features of an applicant to a benefit
    - e.g., a credit request to a bank and the corresponding on credit approval/ denial
  - some designated **categories of applicants**, such as groups protected by anti-discrimination laws,
- find whether, and in which circumstances, there are evidences of discrimination of the designated categories that emerge from the data.

How? Fight with the same weapons

- Idea: use **data mining to discover discrimination**
  - the decision policies hidden in a database can be represented by **decision rules** and discovered by **frequent pattern mining**
  - Once found all such decision rules, highlight all potential **niches of discrimination** by filtering the rules using a measure that quantifies the **discrimination risk**.

German Credit dataset

GERMAN
CHECKING_STATUS
DURATION
CREDIT_HISTORY
PURPOSE
CREDIT_AMOUNT
SAVINGS_STATUS
EMPLOYMENT
INSTALLMENT_COMMITMENT
PERSONAL_STATUS
OTHER_PARTIES
RESIDENCE_SINCE
PROPERTY_MAGNITUDE
AGE
OTHER_PAYMENT_PLANS
HOUSING
EXISTING_CREDITS
JOB
NUM_DEPENDENTS
OWN_TELEPHONE
FOREIGN_WORKER
CREDIT

CHECKING_STATUS	DURATION	CREDIT_HISTORY	PURPOSE	CREDIT_AMOUNT
ge_200	le_17d6	existing_paid	furniture_or_equipment	le_38848d8
no_checking	gt_31d2	existing_paid	radio_or_tv	le_38848d8
no_checking	gt_31d2	existing_paid	used_car	from_7519d6_le_11154d4
no_checking	le_17d6	critical_or_other_existing_credit	radio_or_tv	le_38848d8
lt_0	le_17d6	critical_or_other_existing_credit	other	le_38848d8
from_0_lt_200				
lt_0				
lt_0				
lt_0				
from_0_lt_200				

SAVINGS_STATUS	EMPLOYMENT	INSTALLMENT_COMMITMENT	PERSONAL_STATUS	OTHER_PARTIES
lt_100	lt_1	gt_2d8	female_div_or_dep_or_mar	none
no_known_savings	from_1_lt_4	gt_2d8	female_div_or_dep_or_mar	none
lt_100	from_1_lt_4	le_1d6	female_div_or_dep_or_mar	none
no_known_savings	ge_7	gt_2d8	male_single	none
lt_100	ge_7	gt_2d8	male_single	none

RESIDENCE_SINCE	PROPERTY_MAGNITUDE	AGE	OTHER_PAYMENT_PLANS	HOUSING
le_1d6	life_insurance	from_30d2_le_41d4	bank	own
gt_2d8	car	le_30d2	none	own
from_1d6_le_2d2	life_insurance	le_30d2	bank	own
from_1d6_le_2d2	life_insurance	from_41d4_le_52d6	none	rent
gt_2d8	no_known_property	from_41d4_le_52d6	bank	for_free
le_1d6	real_estate	from_30d2_le_41d4	bank	own
gt_2d8	no_known_property	from_30d2_le_41d4	none	own

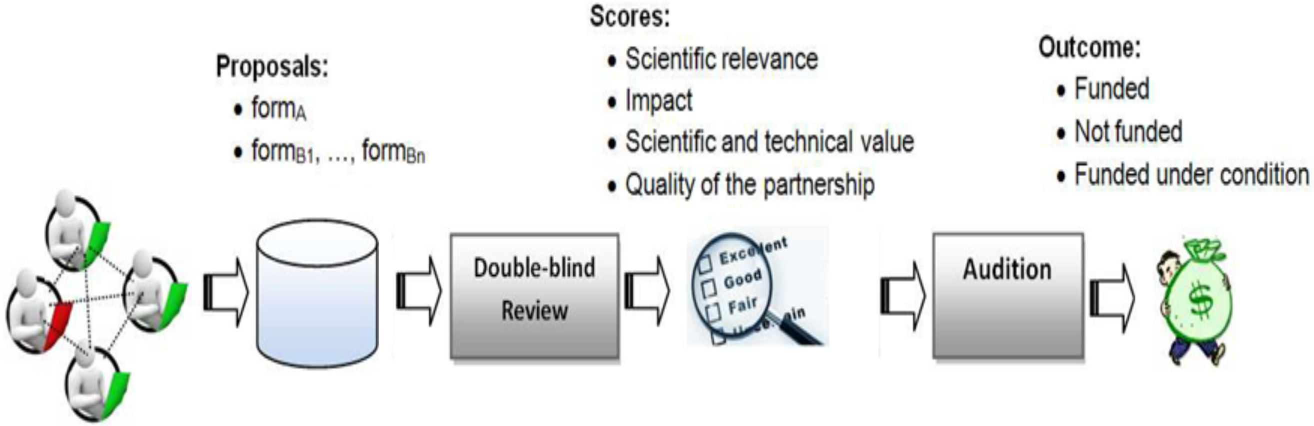
EXISTING_CREDITS	JOB	NUM_DEPENDENTS	OWN_TELEPHONE	FOREIGN_WORKER	CREDIT
le_1d6	high_qualif_or_self_emp_or_mgmt	le_1d2	yes	yes	good
le_1d6	skilled	le_1d2	none	yes	good
le_1d6	skilled	le_1d2	none	yes	good
from_1d6_le_2d2	unskilled_resident	le_1d2	yes	yes	good
from_1d6_le_2d2	high_qualif_or_self_emp_or_mgmt	gt_1d2	yes	yes	good
from_1d6_le_2d2	unskilled_resident	le_1d2	none	yes	good
le_1d6	high_qualif_or_self_emp_or_mgmt	le_1d2	yes	yes	bad
from_1d6_le_2d2	high_qualif_or_self_emp_or_mgmt	le_1d2	none	yes	good
le_1d6	skilled	le_1d2	none	yes	bad

Discrimination discovery from data

- FOREIGN_WORKER=yes
  & PURPOSE=new_car & HOUSING=own
  ➔ CREDIT=bad
  - elift = 5,19 supp = 56 conf = 0,37

elift = 5,19 means that foreign workers have more than 5 times more probability of being refused credit than the average population (even if they own their house).

Case Study: grant evaluation



- Outcome:
  - Funded
  - Not funded
  - Conditionally funded

Dataset attributes

Name	Description	Type	Range/Nominal values	Mean/Mode
Features on the principal and associate investigators				
gender	Gender of principal investigator (PI)	Nominal	{Male, Female}	Male
region	Region of the institution of the PI	Nominal	{North, Center, South}	Center
city	City of the institution of the PI		sta, Aquila, . . . , Trento}	Rome
inst_type	Type of the institution of the PI		iv, Consortium, Other}	Univ
title	Title of the PI		searcher, Prof., Other, PhD}	PhD
age	Age of the PI	Numeric	[26, 39]	32.8
pub_num	Number of publications of the PI	Numeric	[1, 156]	16.4
avg_aut	Average number of authors in publications of the PI	Numeric	[1, 87.1]	4.8
f_partner_num	Number of female principal or associate investigators	Numeric	[0, 3]	0.86
Project costs (absolute values are in €)				
tot_exp	Total cost of the project	Numeric	[300000, 2000000]	971792
fund_req	Requested grant	Numeric	[83720, 1260000]	506205
fund_req_perc	Percentage of requested grant over tota		[26, 63]	51.6
yr_num	Number of young researchers		[1, 10]	2.1
yr_cost	Cost of young researchers		[60000, 981261]	240557
yr_perc	Percentage of young researcher costs over total cost	Numeric	[3, 63]	25.5
grr_num	Number of International good repute researchers	Numeric	[1, 8]	1.5
grr_cost	Cost of good reputation researchers	Numeric	[3500, 610000]	61863
grr_perc	Percentage of good reputation researchers cost	Numeric	[0, 35]	6.1
Research area				
program	Program the project was submitted to		P1, P2}	P2
d1_lv1, d2_lv1, d3_lv1	1^{st}, 2^{nd} and 3^{rd} domain at the 1^{st}		LS, SH, PE}	PE
d1_lv2, d2_lv2, d3_lv2	1^{st}, 2^{nd} and 3^{rd} domain at the 2^{nd}		LS_1, LS_2, . . . , PE_8}	PE_6
d1_lv3, d2_lv3, d3_lv3	1^{st}, 2^{nd} and 3^{rd} domain at the 3^{rd} level of the ERC hierarchy	Nominal	{LS_1_1, LS_1_2, . . . , PE_8_15}	PE_6_17
Project evaluation				
s1	Scores S1 received at the peer-review	Numeric	[1, 8]	6.6
s2	Scores S2 received at the peer-review	Numeric	[1, 7]	5.7
s3	Scores S3 received at the peer-revie			11.8
s4	Scores S4 received at the peer-revie			8.1
audition	Whether the project passed the peer-review		{yes, no}	no
funded	Whether the project was funded (2nd evaluation phase)	Nominal	{yes, no, conditionally}	no
fund	The actual granted amount after budget cut	Numeric	[228000, 750100]	429990

Features of the PI

Project costs

Research Area

Project Evaluation

A potentially discriminatory rule

```
R2: (d1_lv2 = PE4) and (tot_cost >= 1,358,000) and
    (age <= 35) => disc=yes
    [prec=1.0] [rec=0.031] [diff=0.194] [OR=4.50]
```

- Antecedent
  - Project proposals in “Physical and Analytical Chemical Sciences”
  - Young females
  - Total cost of 1,358,000 Euros or above
- Possible interpretation
  - *“Peer-reviewers of panel PE4 trusted young females requiring high budgets less than males leading similar projects”*

Case study: US Harmonized Tariff System

- US Harmonized Tariff System (HTS)
- https://hts.usitc.gov/
- ***Detailed** tariff classification system for merchandise imported to US*
- Chapter 61, 62, 64, 65: apparels
  - Different taxes for same garments separately p luced for male and emale
  - D uctured

Heading/ Subheading	Stat. Suf-fix	Article Description	Unit of Quantity	Rates of Duty 1 General	Rates of Duty 1 Special	Rates of Duty 2
6112 (con.)		Track suits, ski-suits and swimwear, knitted or crocheted (con.):				
		Men's or boys' swimwear:				
6112.31.00		Of synthetic fibers		25.9%	Free (BH,CA, CL,IL,MX, P,SG) 5.3% (JO) 7.8% (MA) 15.5% (AU)	90%
	10	Men's (659)	doz. kg			
	20	Boys' (659)	doz. kg			
6112.39.00		Of other textile materials		13.2%	Free (BH,CA, CL,E*,IL,JO, MX,P,SG) 3.9% (MA) 11.8% (AU)	90%
	10	Of cotton (359)	doz. kg			
		Other:				
	15	Containing 70 percent or more by weight of silk or silk waste (759)	doz. kg			
	90	Other (859)	doz. kg			
		Women's or girls' swimwear:				
6112.41.00		Of synthetic fibers		24.9%	Free (BH,CA, CL,IL,MX, P,SG) 5.1% (JO) 7.5% (MA)	90%

Women: 14%
Men: 9%

1.3 billions USD!!!

The New York Times

BUSINESS DAY

In Apparel, All Tariffs Aren't Created Equal

By MICHAEL BARBARO APRIL 28, 2007

Totes-Isotoner Corp. v. U.S.

Rack Room Shoes Inc. and Forever 21 Inc. vs U.S.

Court of International Trade

U.S. Court of Appeals for the Federal Circuit (2014)

"[…] the courts may have concluded that Congress had no discriminatory intent when ruling the HTS, but there is little doubt that gender-based tariffs have **discriminatory impact**"

Fairer Trade

Removing Gender Bias in US Import Taxes

LORI L. TAYLOR AND JAWAD DAR
Mosbacher Institute

There are many inequalities in US tariff policy. Products imported from certain countries enter duty free, while nearly identical products from other countries are heavily taxed. Tariffs on agricultural products are systematically higher than those on manufactured goods. Tariffs on some categories of manufactured goods—such as shoes or cotton shirts—depend on the gender of the intended consumer. Some of these tariff differences have a rational basis in the policy interests of the United States. However, differential taxation of apparel based on gender cannot be defended and should be abolished.

Sample rule from the HTS dataset

$$Shorts(?x) \wedge hasMaterial(?x, \text{“}fine\ \ animal\ \ hair\text{”}) \\ \quad \rightarrow isDiscriminatory(?x, yes)$$

with a confidence $conf = 66.67\%$ can be directly compared with its ancestor rule at the grand-parent level (the concept $Shorts$ is a sub-class of $Outerwear$):

$$Outerwear(?x) \wedge hasMaterial(?x, \text{“}fine\ \ animal\ \ hair\text{”}) \\ \quad \rightarrow isDiscriminatory(?x, yes)$$

which has a lower confidence of $conf = 57.78\%$.

Explaining human decision making

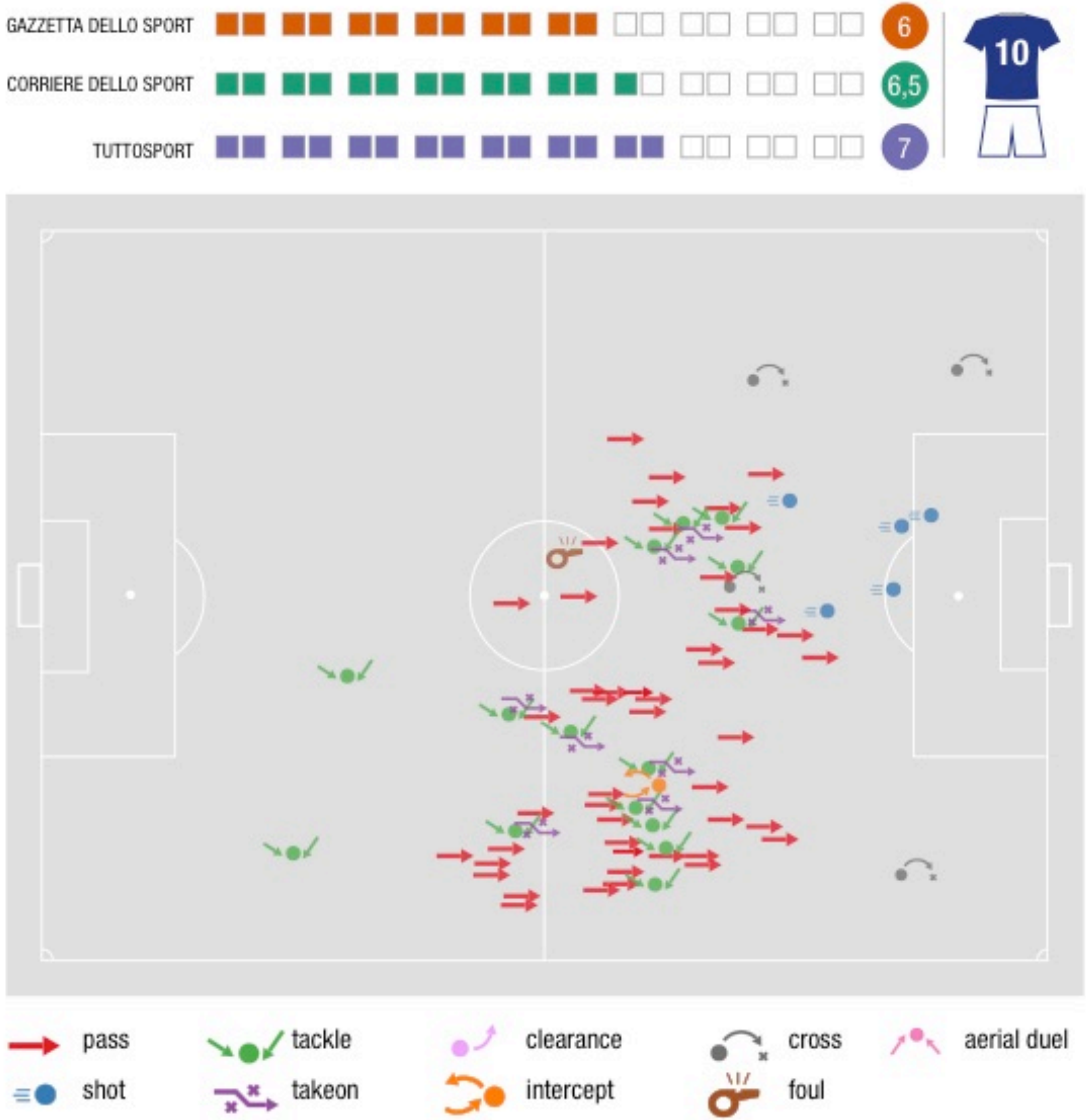
L. Pappalardo, P. Cintia, F. Giannotti, D. Pedreschi, A.-L. Barabasi.

The human perception of performance (forthcoming)

Soccer Player Ratings

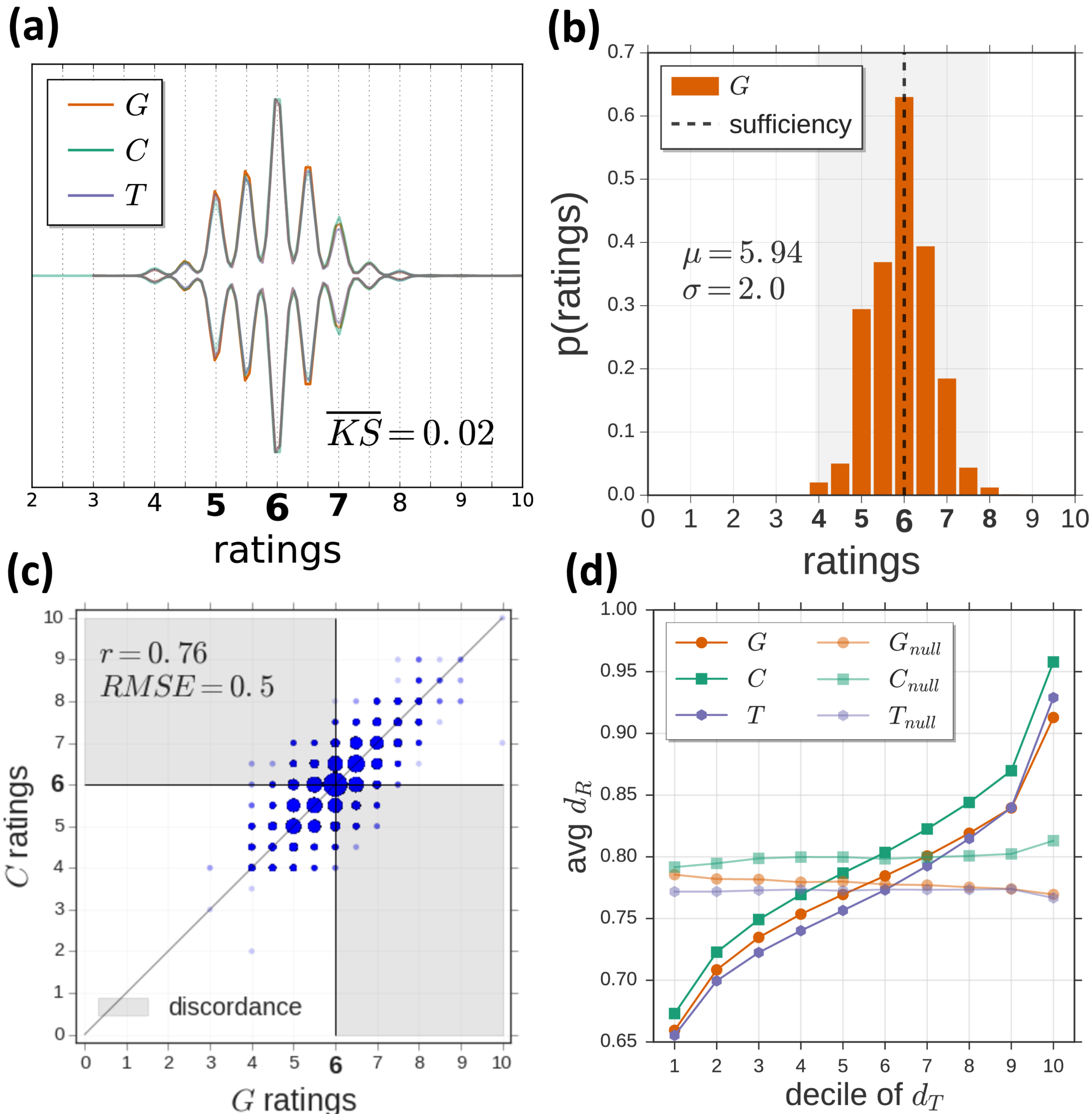
Soccer Player Ratings

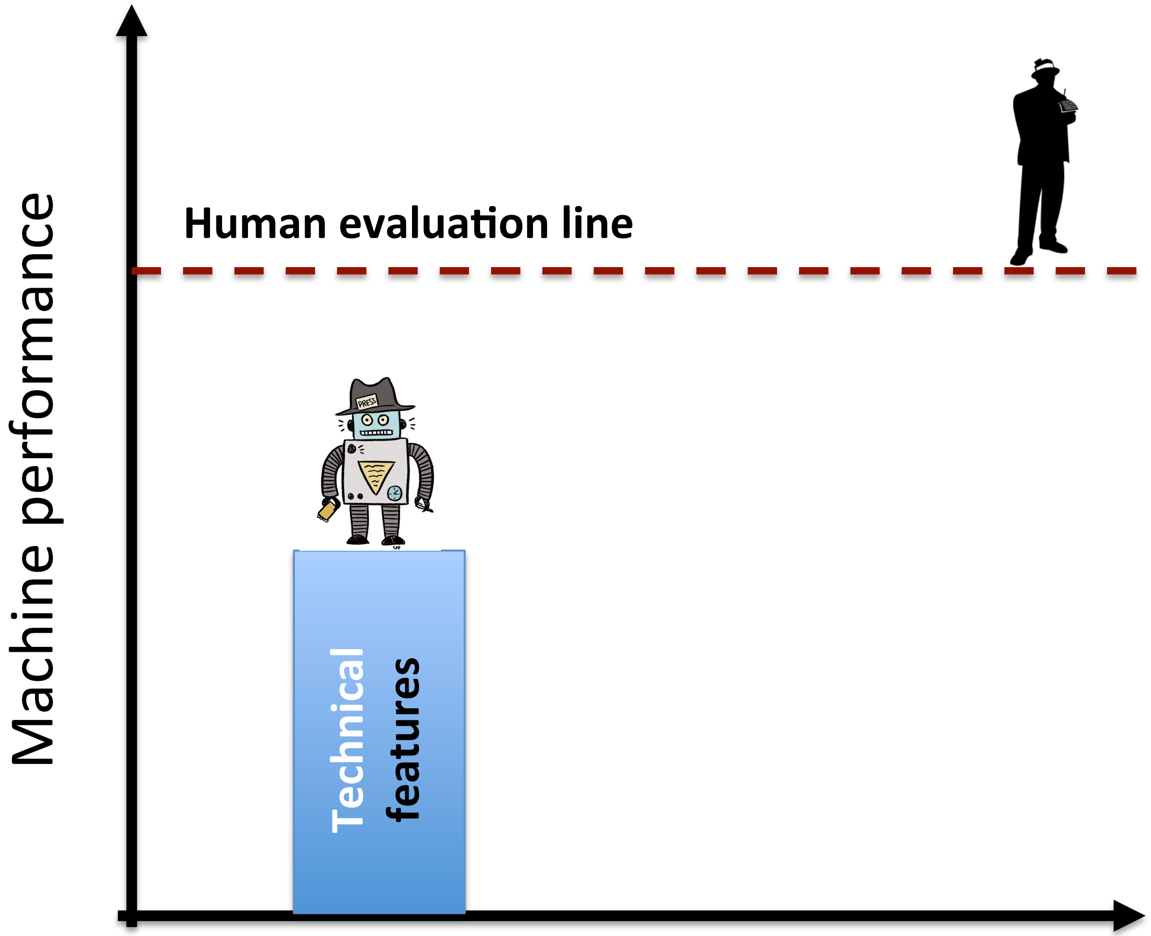
How humans evaluate sports performance?

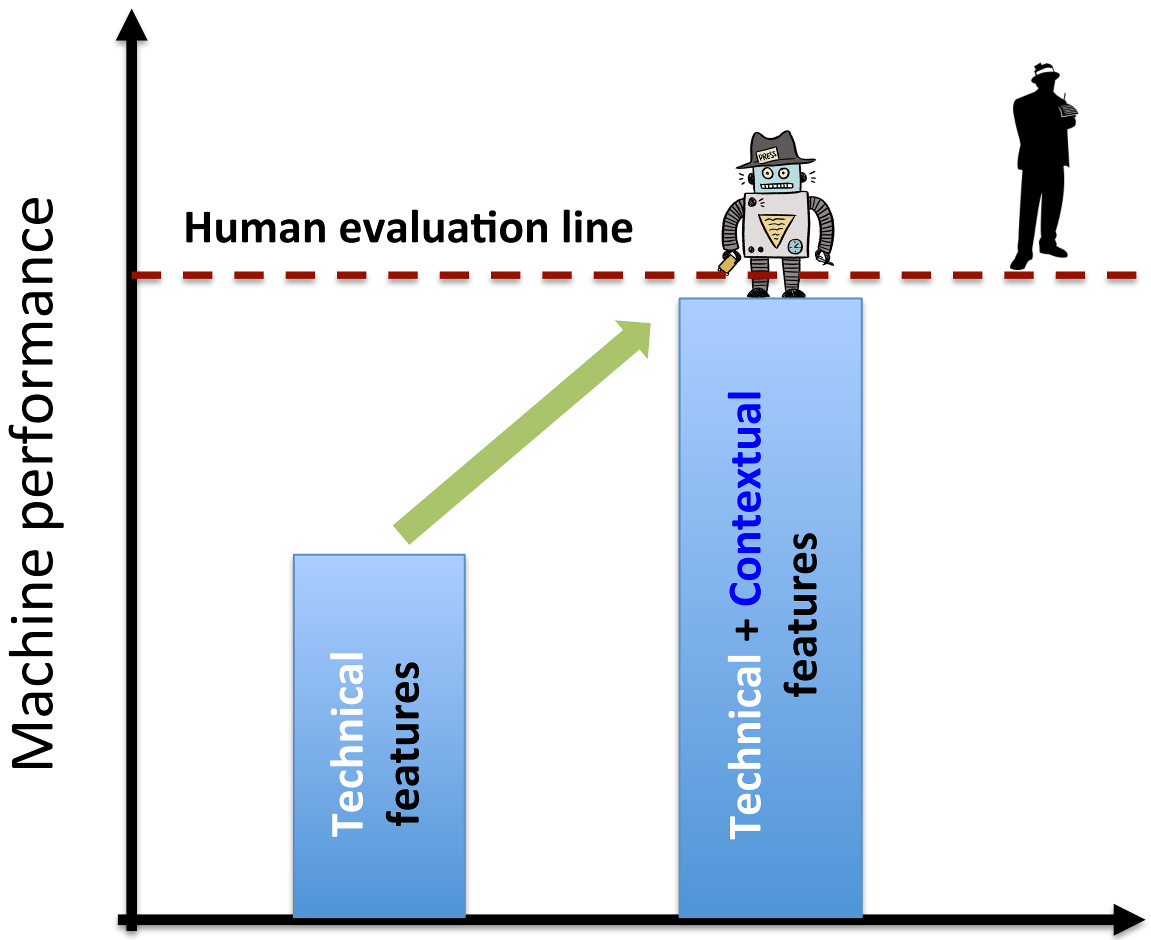

GAZZETTA DELLO SPORT
6
CORRIERE DELLO SPORT
6,5
TUTTOSPORT
7
10
pass
shot
tackle
takeon
clearance
intercept
cross
foul
aerial duel

Observe, predict, explain

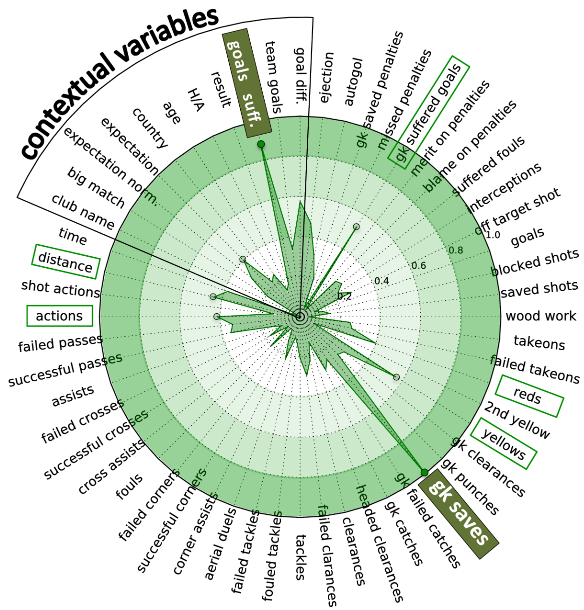
- Observe
  - Use sensed data to measure and quantify different aspects of human performance, together with associated score
- Predict
  - Construct a predictive model using machine learning from data
- Explain
  - Explain the obtained model to discover rules adopted by the model to score a performance, thus reproducing the logic of the human evaluator


(a)
G
C
T
$\overline{KS} = 0.02$
2 3 4 5 6 7 8 9 10
ratings
(b)
G
sufficiency
$\mu = 5.94$
$\sigma = 2.0$
p(ratings)
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
0 1 2 3 4 5 6 7 8 9 10
ratings
(c)
$r = 0.76$
$RMSE = 0.5$
discordance
C ratings
G ratings
0 1 2 3 4 5 6 7 8 9 10
(d)
G
C
T
G_{null}
C_{null}
T_{null}
avg d_R
0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
1 2 3 4 5 6 7 8 9 10
decile of d_T

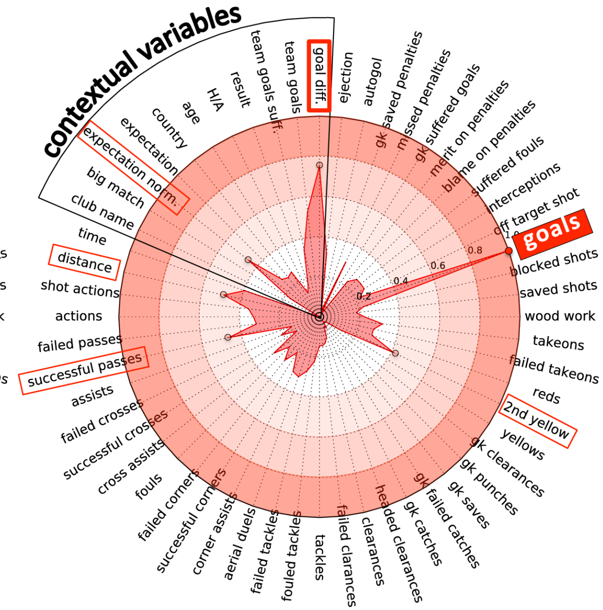

Human evaluation line
Machine performance
Technical features


Machine performance
Human evaluation line
Technical features
Technical + Contextual features

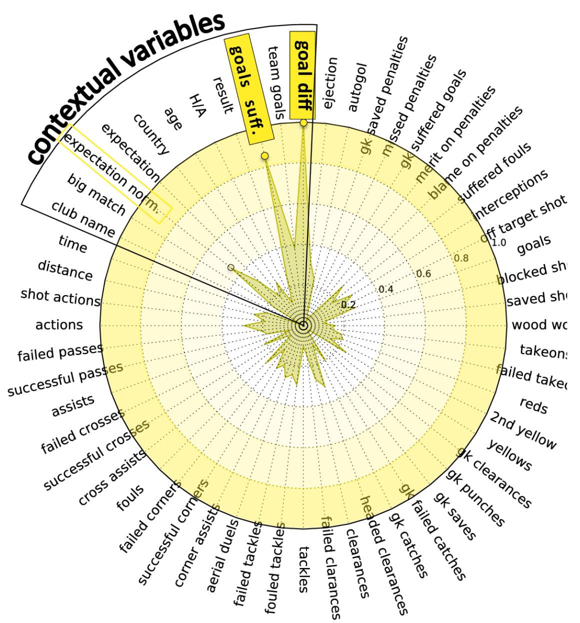
Goalkeepers



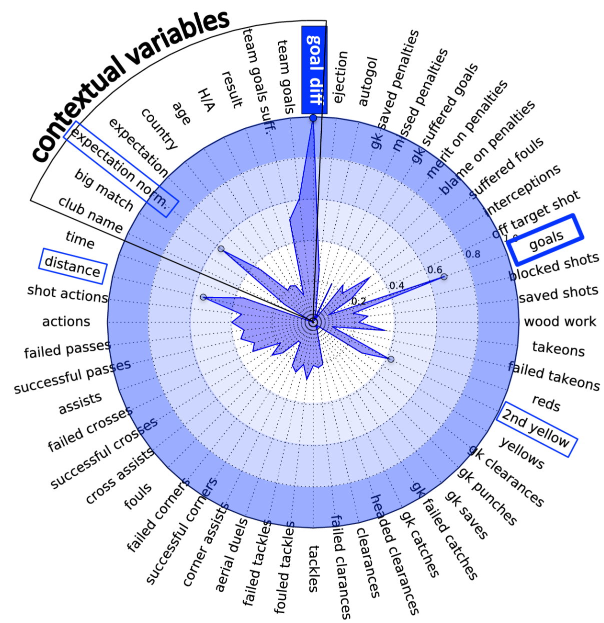
Forwards



Defenders



Midfielders



SoBigData Research Infrastructure

Social Mining & Big Data Analytics

H2020 - www.sobigdata.eu

September 2015- August 2019

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

IIT
Istituto di Informatica e Telematica del CNR

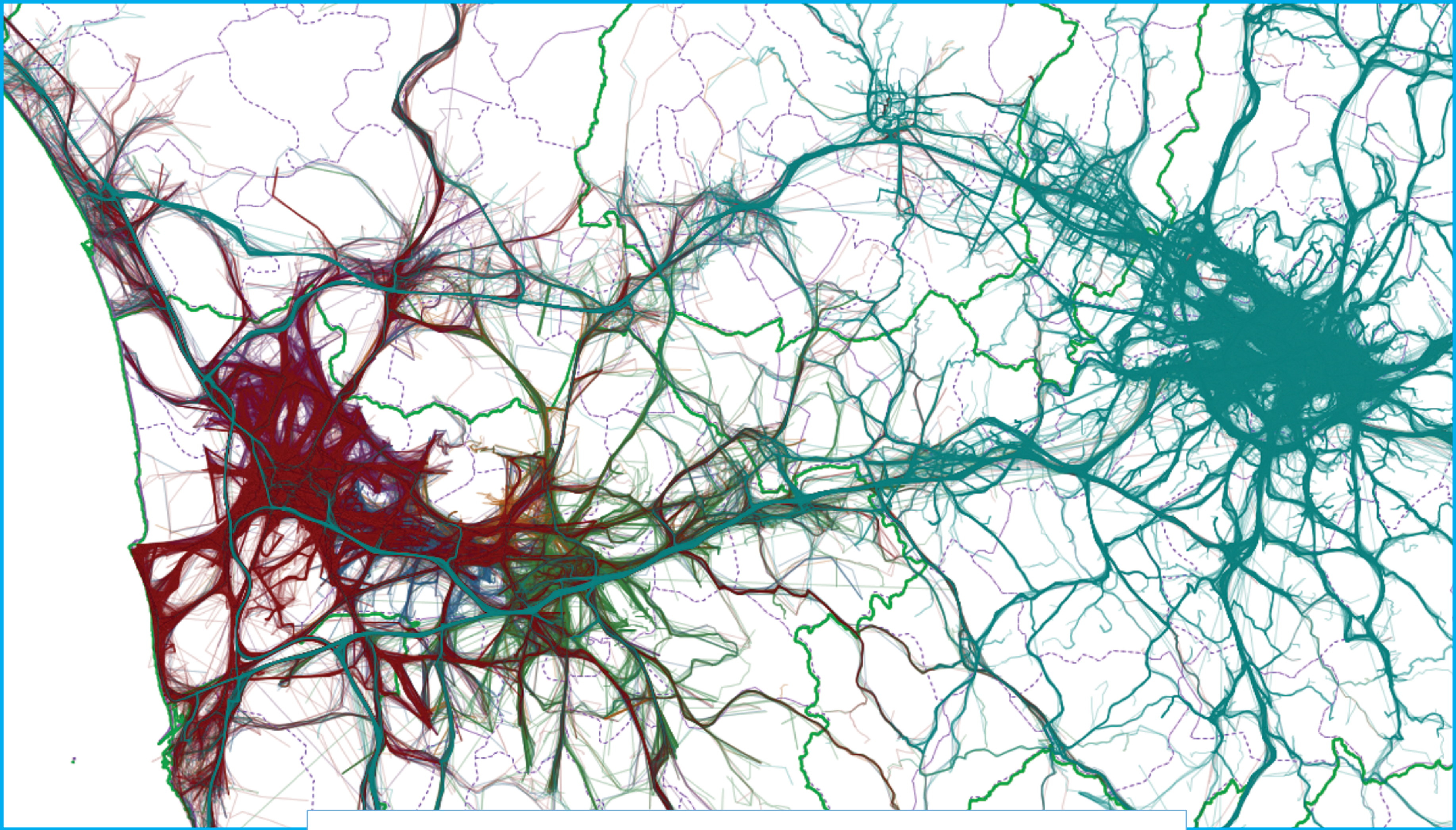
IN SUPREMAE DIGNITATIS 1343
UNIVERSITÀ DI PISA

SCUOLA NORMALE SUPERIORE

Institutions Markets Technologies
IMT
INSTITUTE FOR ADVANCED STUDIES LUCCA

Exploratory:

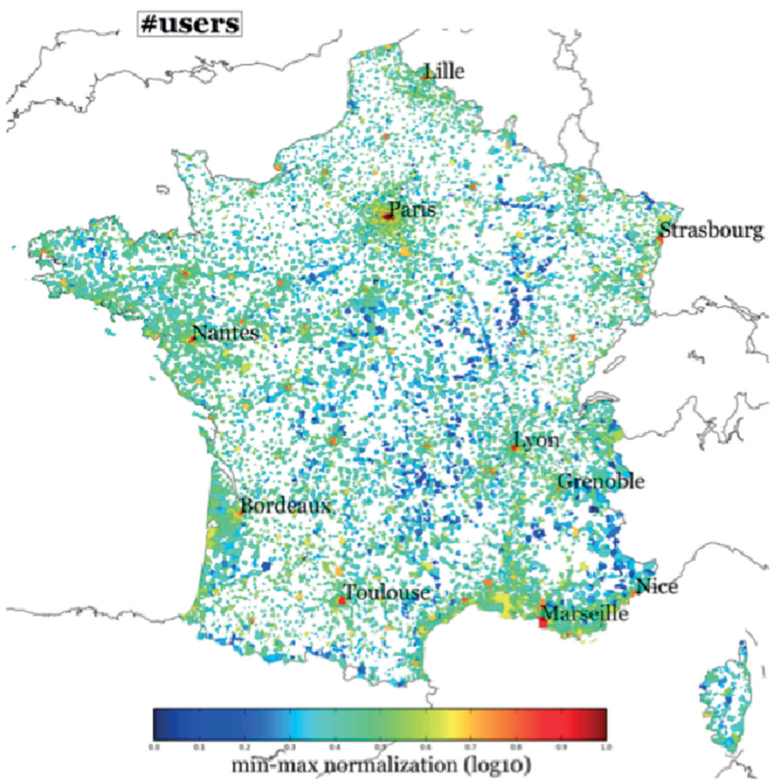
Big Data for City of Citizens

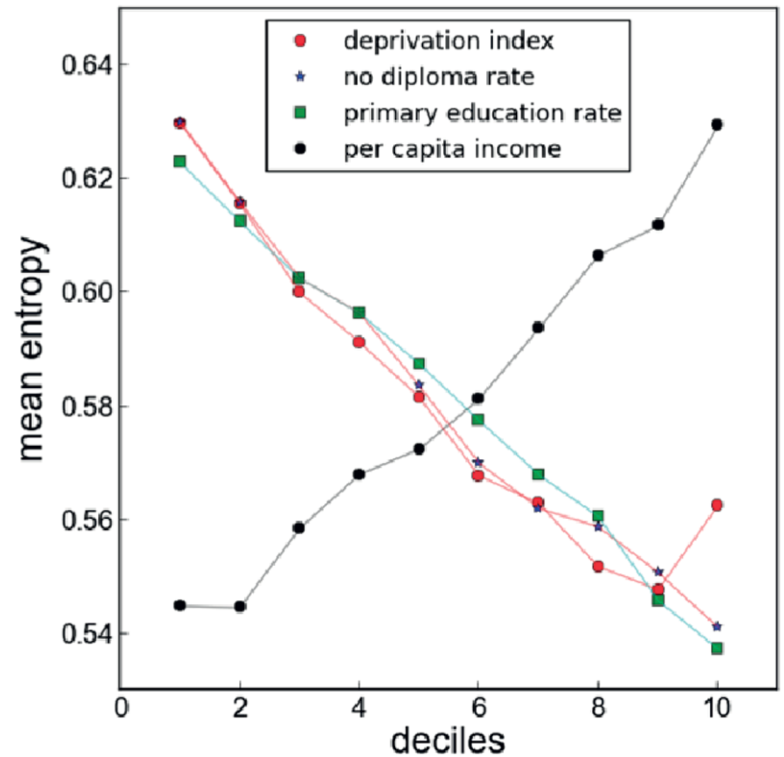


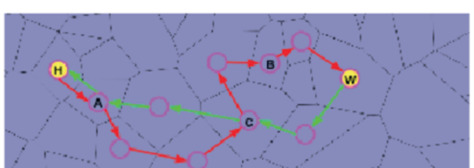
Personal Mobility, Social + Mobility, Personal Sensing

Exploratory:

Big Data for Well Being and Economic Performance







$$d_i^{(n)} = \sum_{j=1}^{|V|} \frac{1}{k_j} M_{ij} p_j^{(n-1)} \forall i \qquad p_j^{(n)} = \sum_{i=1}^{|U|} \frac{1}{k_i} M_{ij} d_i^{(n-1)} \forall j$$

Deprivation Index (in France) predicted with Mobile Phone traces

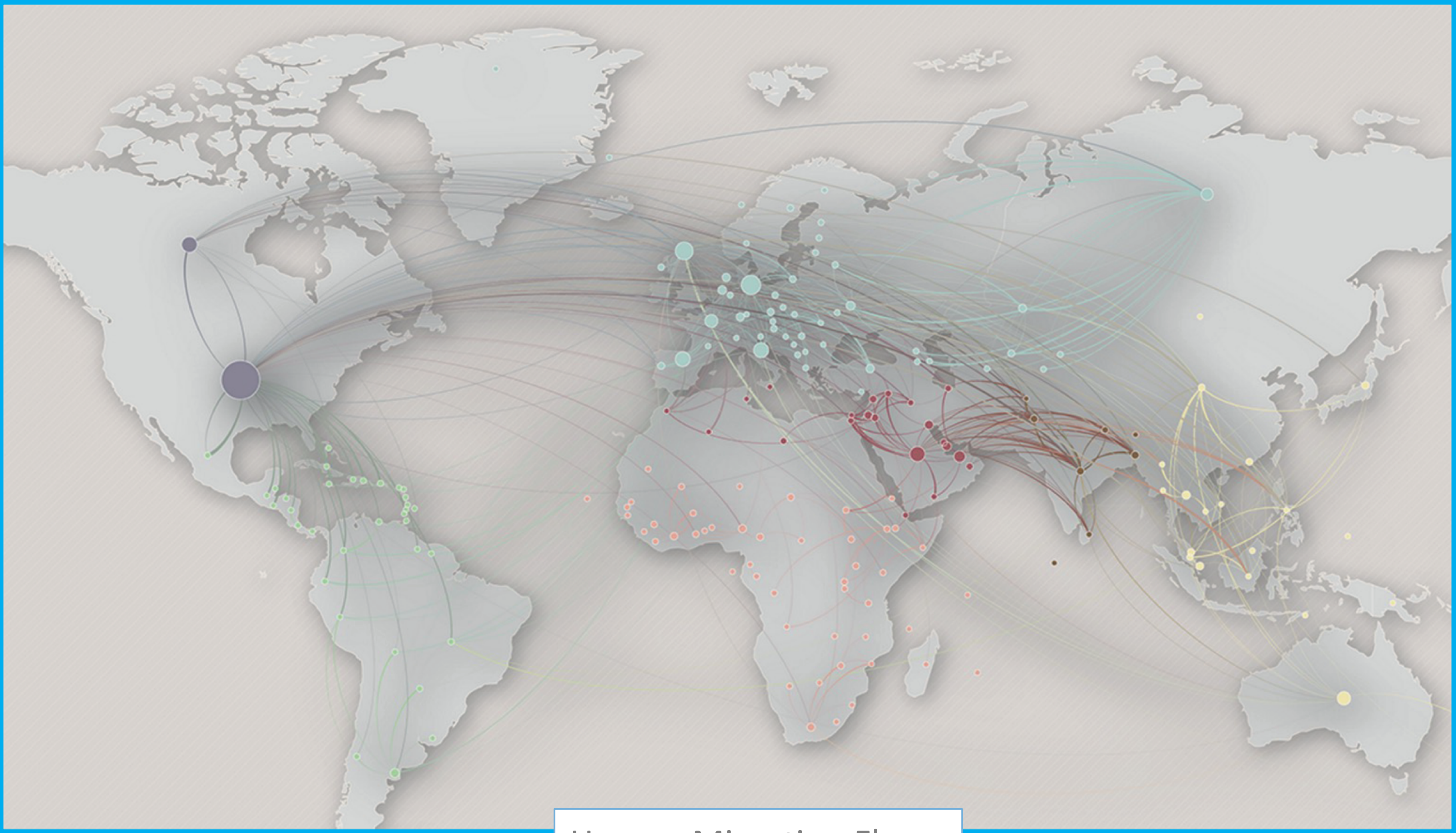
Exploratory:

Big Data for Societal Debates

Polarization, controversy and topic trends on societal debates through social media

Next Exploratory:

Big Data for Migration Studies



Human Migration Flows

Ethics and Security

Legal and Ethical framework

Define and implement the legal and ethical framework of the SoBigData RI, in accordance with the European and national legislations

Monitor of research

Monitor the compliance of experiments and research protocols with the framework

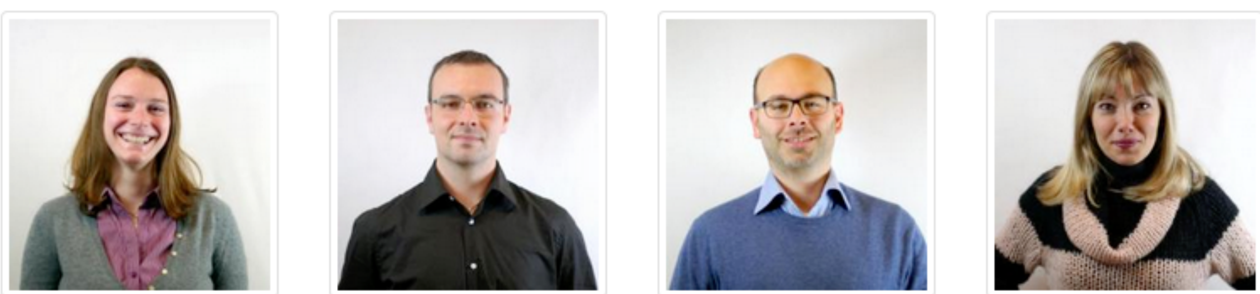
Privacy-by-design

The development of big data analytics and social mining tools with Value-Sensitive Design and privacy-by-design methodologies

This is the work of many people for a long time

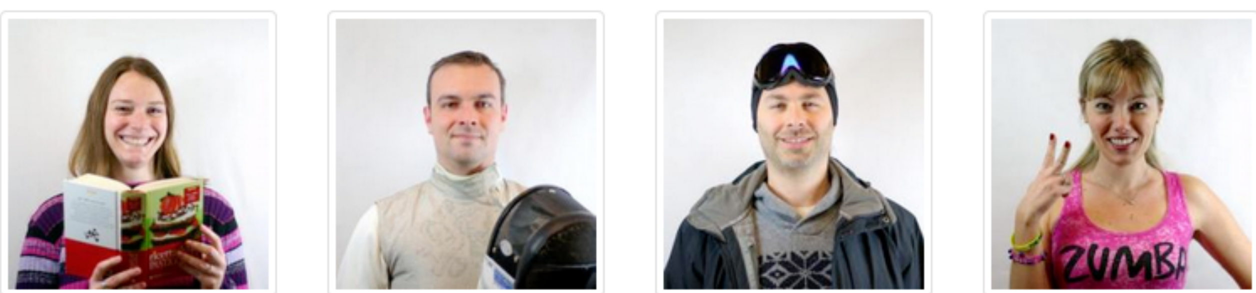
- Fosca Giannotti, Mirco Nanni, Salvo Rinzivillo, Roberto Trasarti, Anna Monreale, Salvatore Ruggieri, Franco Turini
- all the fantastic folks at KDD LAB Pisa
  - http://kdd.isti.cnr.it

- Many international collaborators
- Thanks a lot!



Knowledge Discovery & Data Mining Lab

http://kdd.isti.cnr.it



Knowledge Discovery & Data Mining Lab

http://kdd.isti.cnr.it

Key publications

- F Giannotti, M Nanni, F Pinelli, D Pedreschi. Trajectory pattern mining. ACM SIGKDD 2007
- F Giannotti, D Pedreschi. Mobility, data mining and privacy: Geographic knowledge discovery. Springer, 2008
- A Monreale, F Pinelli, R Trasarti, F Giannotti. WhereNext: a location predictor on trajectory pattern mining. ACM SIGKDD 2009
- S Rinzivillo, D Pedreschi, M Nanni, F Giannotti, N Andrienko, G Andrienko. Visually driven analysis of movement data by progressive clustering. Information Visualization 7 (3-4), 225-239. 2008
- D Wang, D Pedreschi, C Song, F Giannotti, AL Barabasi. Human mobility, social ties, and link prediction. ACM SIGKDD 2011
- F Giannotti, M Nanni, D Pedreschi, F Pinelli, C Renso, S Rinzivillo, R Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. The VLDB Journal 20(5) 2011
- R Trasarti, F Pinelli, M Nanni, F Giannotti. Mining mobility user profiles for car pooling. ACM SIGKDD 2011
- M Coscia, G Rossetti, F Giannotti, D Pedreschi. Demon: a local-first discovery method for overlapping communities. ACM SIGKDD 2012
- D Pennacchioli, M Coscia, S Rinzivillo, F Giannotti, D Pedreschi. The retail market as a complex system. EPJ Data Science 3 (1), 1-27 (2014)
- A Monreale, S Rinzivillo, F Pratesi, F Giannotti, D Pedreschi. Privacy-by-design in big data analytics and social mining. EPJ Data Science 3 (1), 1-26 (2014)
- Luca Pappalardo, Filippo Simini, Salvatore Rinzivillo, Dino Pedreschi, Fosca Giannotti & Albert-László Barabási. Returners and explorers dichotomy in human mobility. Nature Communications 6, Article number: 8166 (2015) doi:10.1038/ncomms9166 (2015)

Key publications

- M Coscia, G Rossetti, F Giannotti, D Pedreschi. Demon: a local-first discovery method for overlapping communities. ACM SIGKDD 2012
- S Rinzivillo, S Mainardi, F Pezzoni, M Coscia, D Pedreschi, F Giannotti. Discovering the geographical borders of human mobility. KI-Künstliche Intelligenz 26 (3) 2012
- D Pennacchioli, M Coscia, S Rinzivillo, D Pedreschi, F Giannotti. Explaining the Product Range Effect in Purchase Data. IEEE BIGDATA 2013
- B Furletti, L Gabrielli, C Renso, S Rinzivillo. Analysis of GSM Calls Data for Understanding User Mobility Behavior. IEEE BIGDATA 2013
- L Milli, A Monreale, G Rossetti, D Pedreschi, F Giannotti, F Sebastiani. Quantification trees. IEEE ICDM 2013
- Giusti, Marchetti, Pratesi, Salvati, Pedreschi, Giannotti, Rinzivillo, Pappalardo, Gabrielli. Small area model based estimators using Big Data Sources. Journal of Official Statistics, 31(2) 2015.
- Furletti, Gabrielli, Garofalo, Giannotti, Milli, Nanni, Pedreschi, 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.
- Luca Pappalardo, Maarten Vanhoof, Zbigniew Smoreda, Dino Pedreschi,Fosca Giannotti. Human Mobility and Economic Development. IEEE BIG DATA (2015).

Vision papers

- F Giannotti, D Pedreschi, A Pentland, P Lukowicz, D Kossmann, J Crowley, D Helbing. **A planetary nervous system for social mining and collective awareness**. The European Physical Journal Special Topics 214 (1), 49-75, 2012
- J van den Hoven, D Helbing, D Pedreschi, J Domingo-Ferrer, F Giannotti . **FuturICT—The road towards ethical ICT**. The European Physical Journal Special Topics 214 (1), 153-181, 2012
- M Batty, KW Axhausen, F Giannotti, A Pozdnoukhov, A Bazzani, M Wachowicz. **Smart cities of the future.** The European Physical Journal Special Topics 214 (1), 481-518, 2012