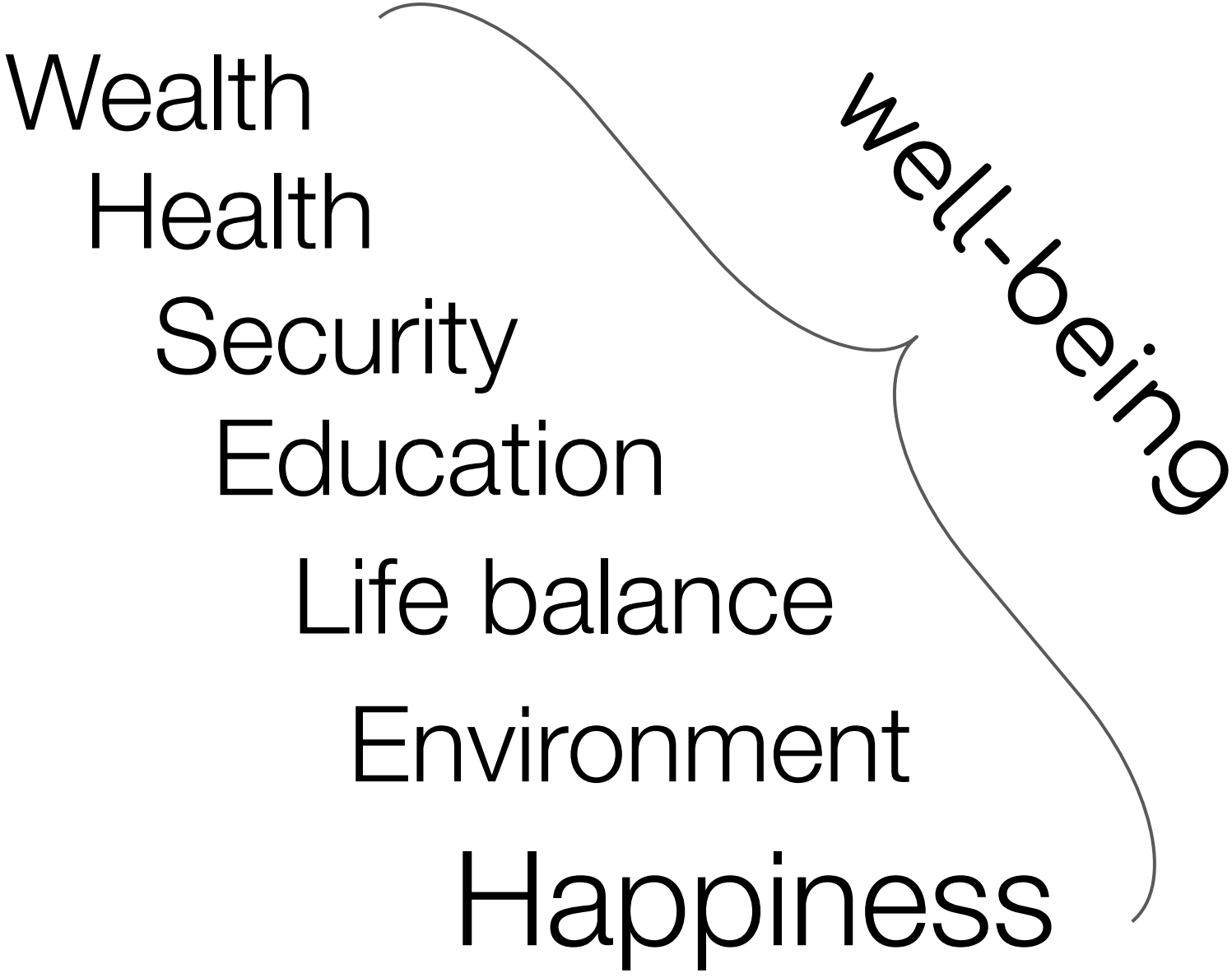
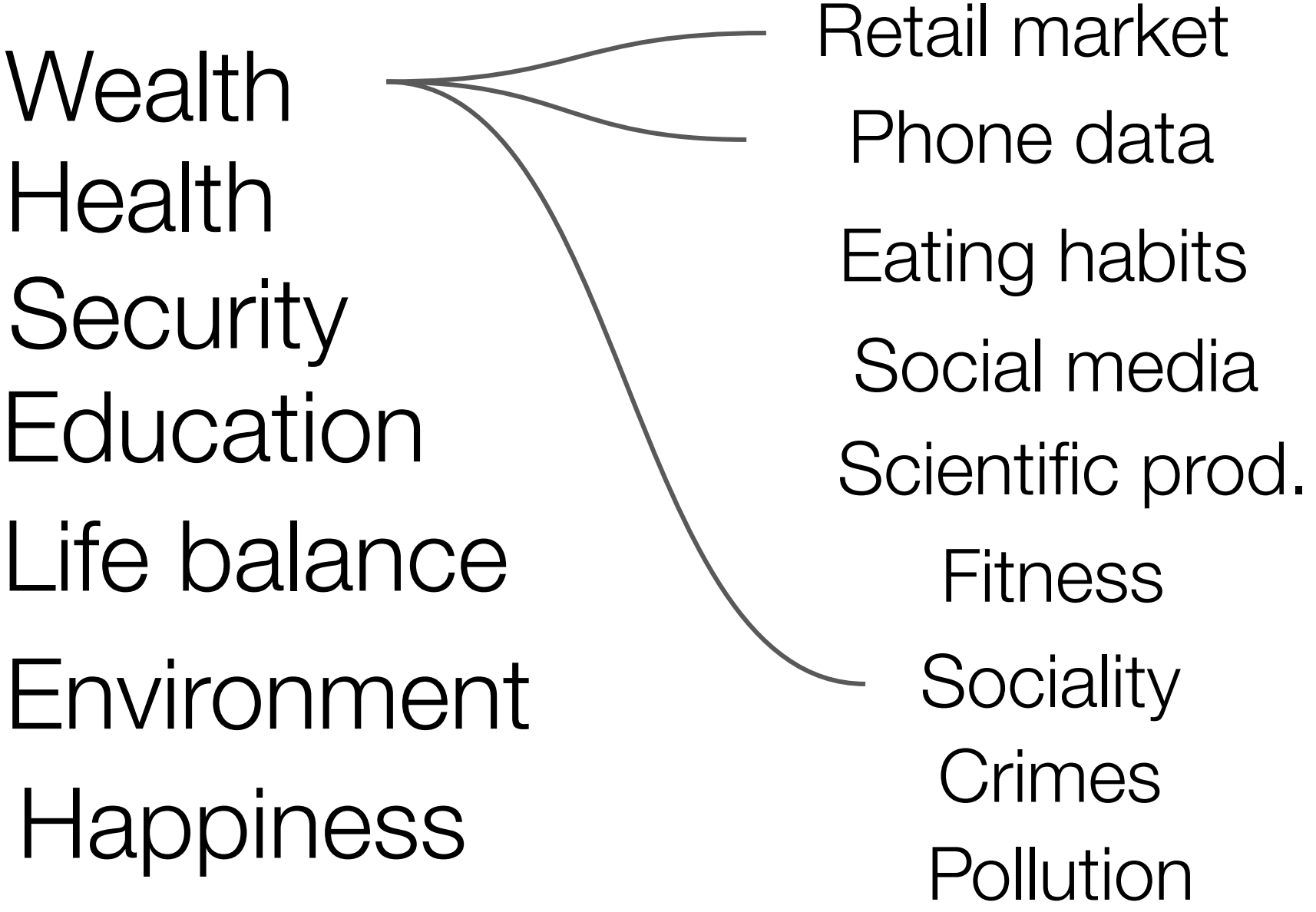
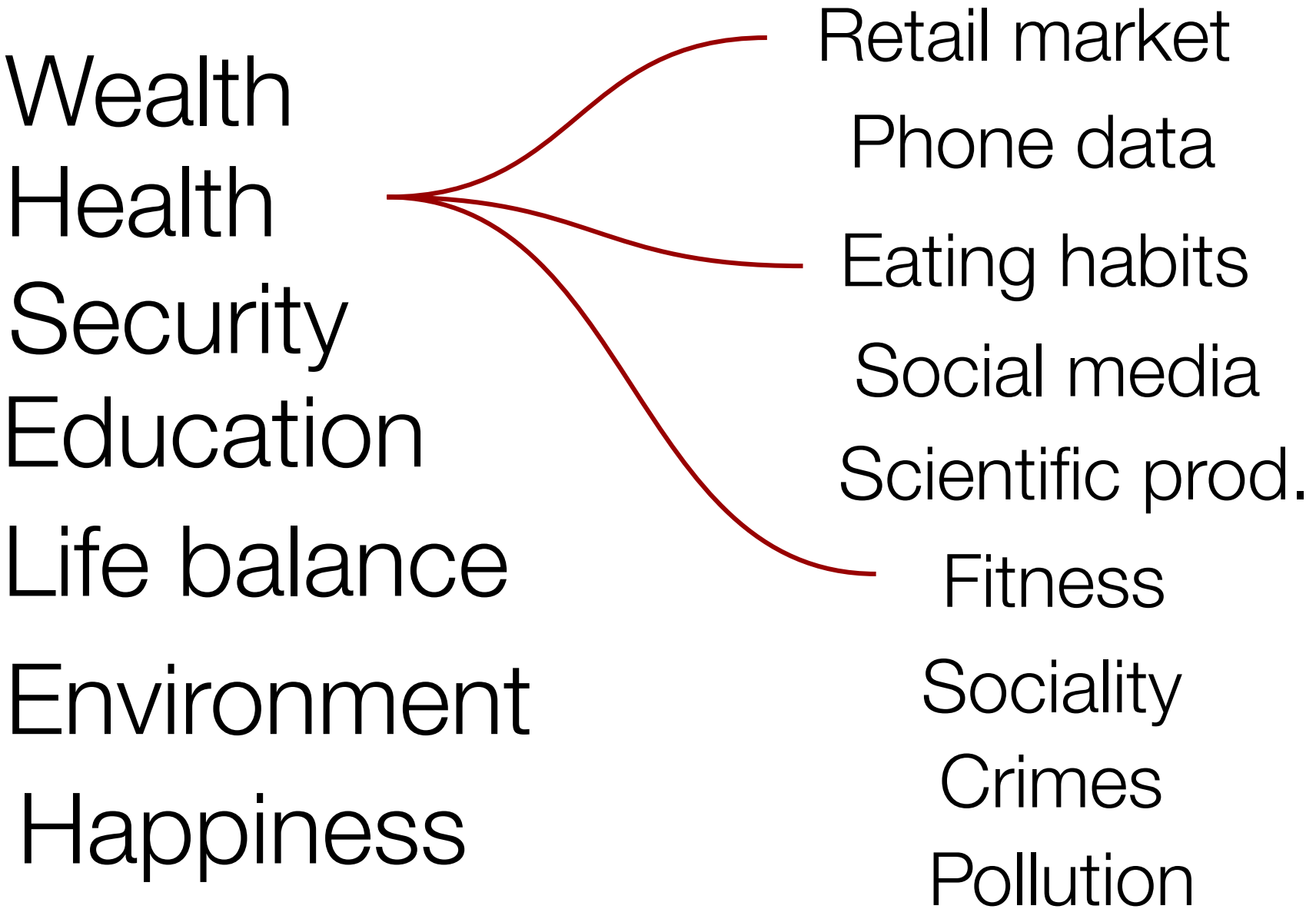
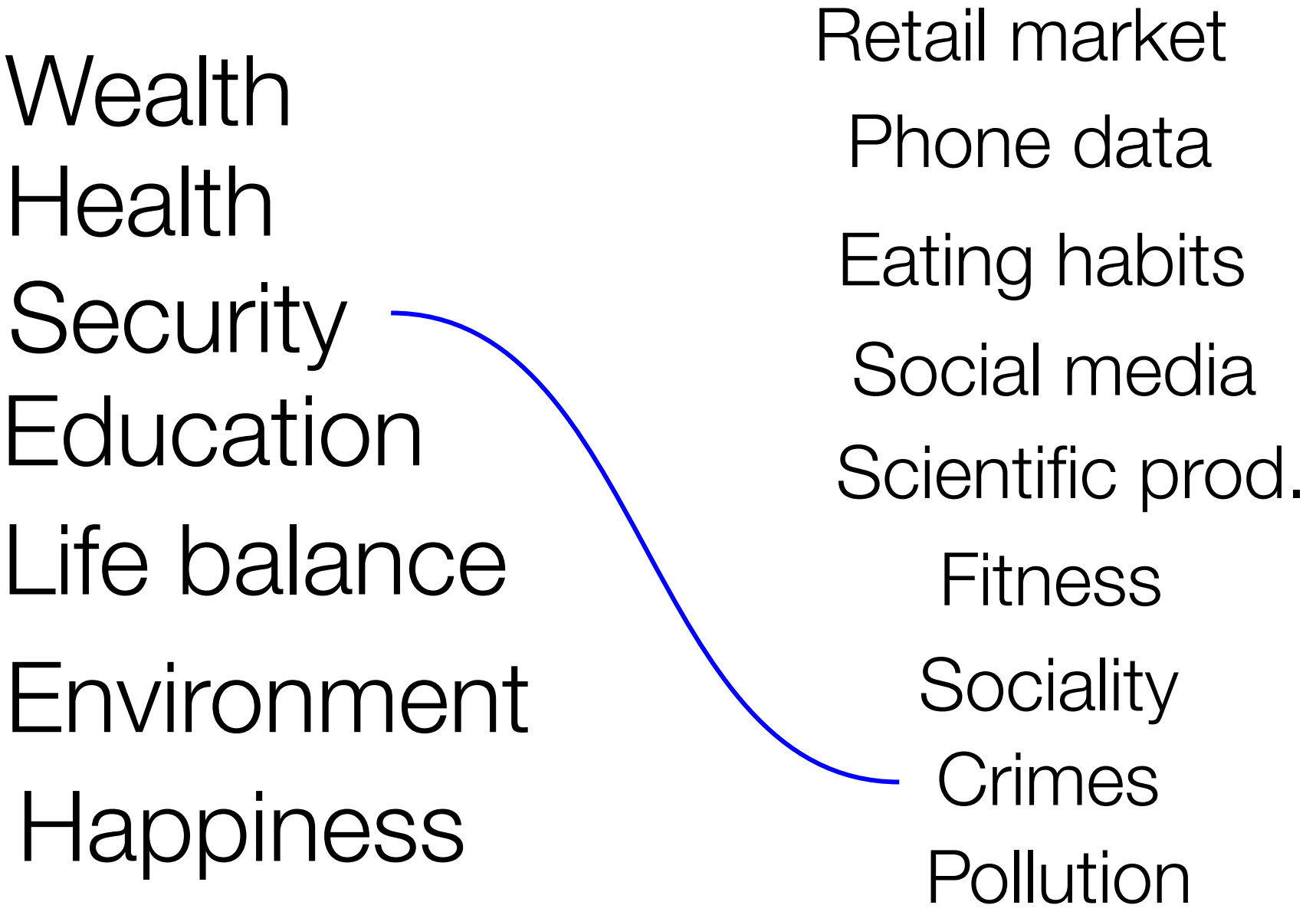
How to use #DataScience and #BigData to nowcast well-being?

How to use #DataScience and #BigData to nowcast **well-being**?

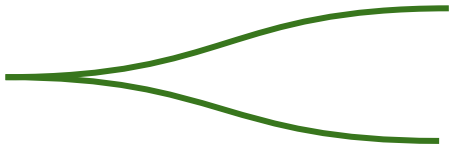

Wealth
Health
Security
Education
Life balance
Environment
Happiness
well-being


Wealth
Health
Security
Education
Life balance
Environment
Happiness
Retail market
Phone data
Eating habits
Social media
Scientific prod.
Fitness
Sociality
Crimes
Pollution


Wealth
Health
Security
Education
Life balance
Environment
Happiness
Retail market
Phone data
Eating habits
Social media
Scientific prod.
Fitness
Sociality
Crimes
Pollution

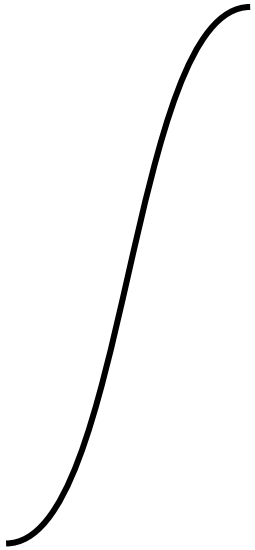

Wealth
Health
Security
Education
Life balance
Environment
Happiness
Retail market
Phone data
Eating habits
Social media
Scientific prod.
Fitness
Sociality
Crimes
Pollution

Wealth
Health
Security
Education
Life balance
Environment
Happiness



Retail market
Phone data
Eating habits
Social media
Scientific prod.
Fitness
Sociality
Crimes
Pollution

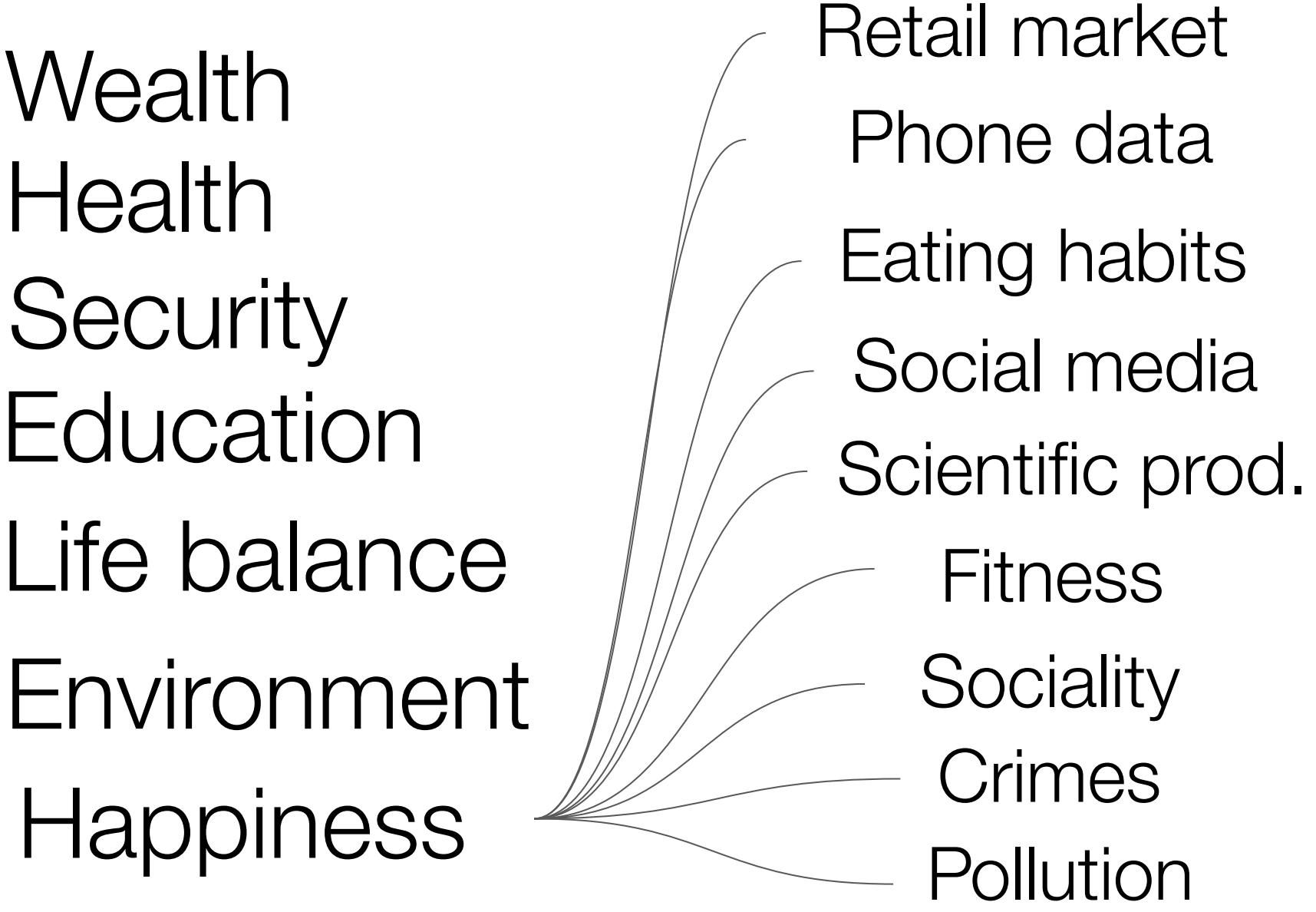
Wealth
Health
Security
Education
Life balance
Environment
Happiness



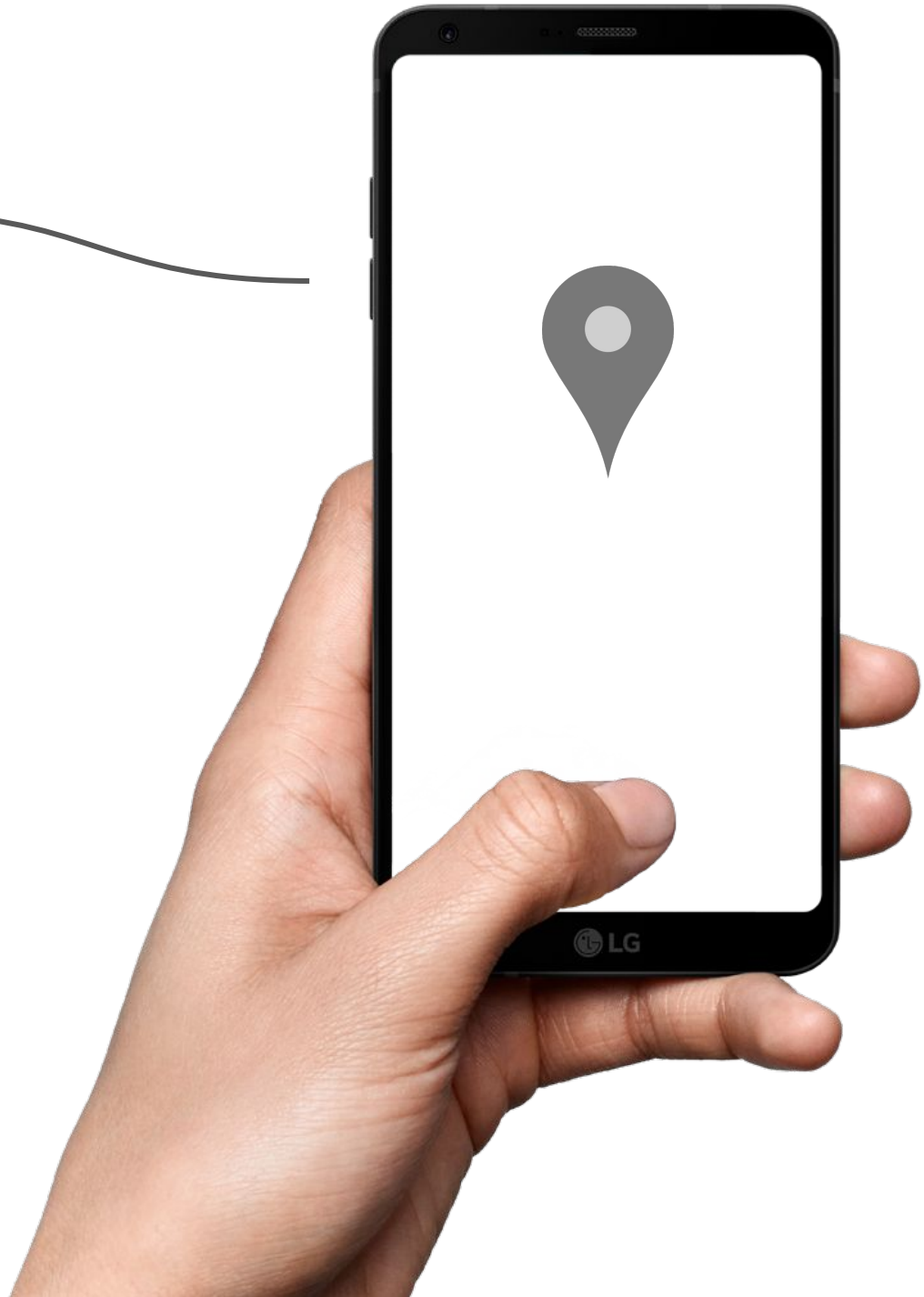
Retail market
Phone data
Eating habits
Social media
Scientific prod.
Fitness
Sociality
Crimes
Pollution

Wealth
Health
Security
Education
Life balance
Environment
Happiness

Retail market
Phone data
Eating habits
Social media
Scientific prod.
Fitness
Sociality
Crimes
Pollution


Wealth
Health
Security
Education
Life balance
Environment
Happiness
Retail market
Phone data
Eating habits
Social media
Scientific prod.
Fitness
Sociality
Crimes
Pollution

Wealth



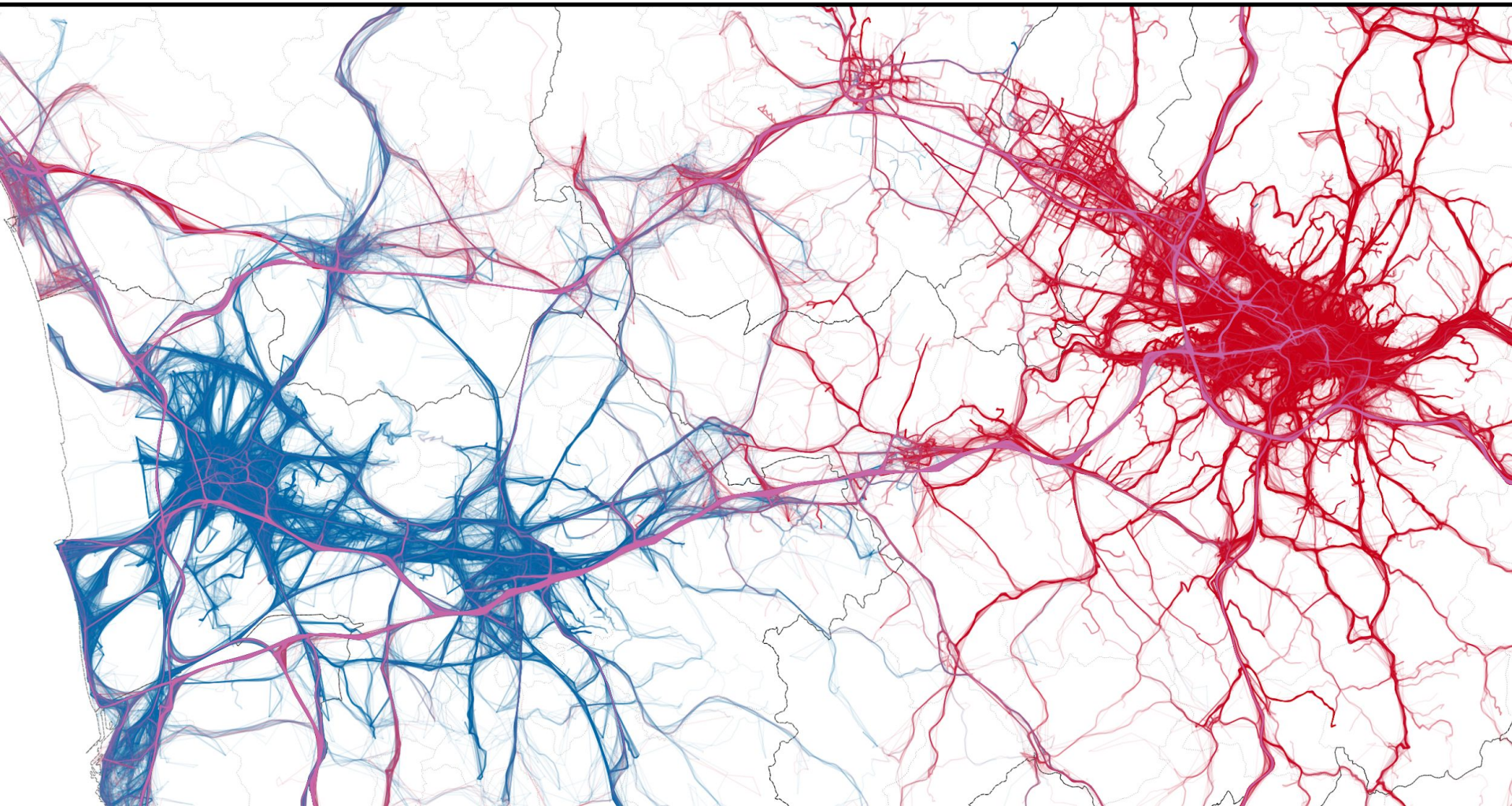
Surveys have limits

1. They do not scale to small territories and they are not dynamic

2. They do not adapt to Pareto distributions:
   – We cannote use samples
   – 50 people are as wealthy as 3.5 billion people

3. Young people do not respond to surveys

Big Data

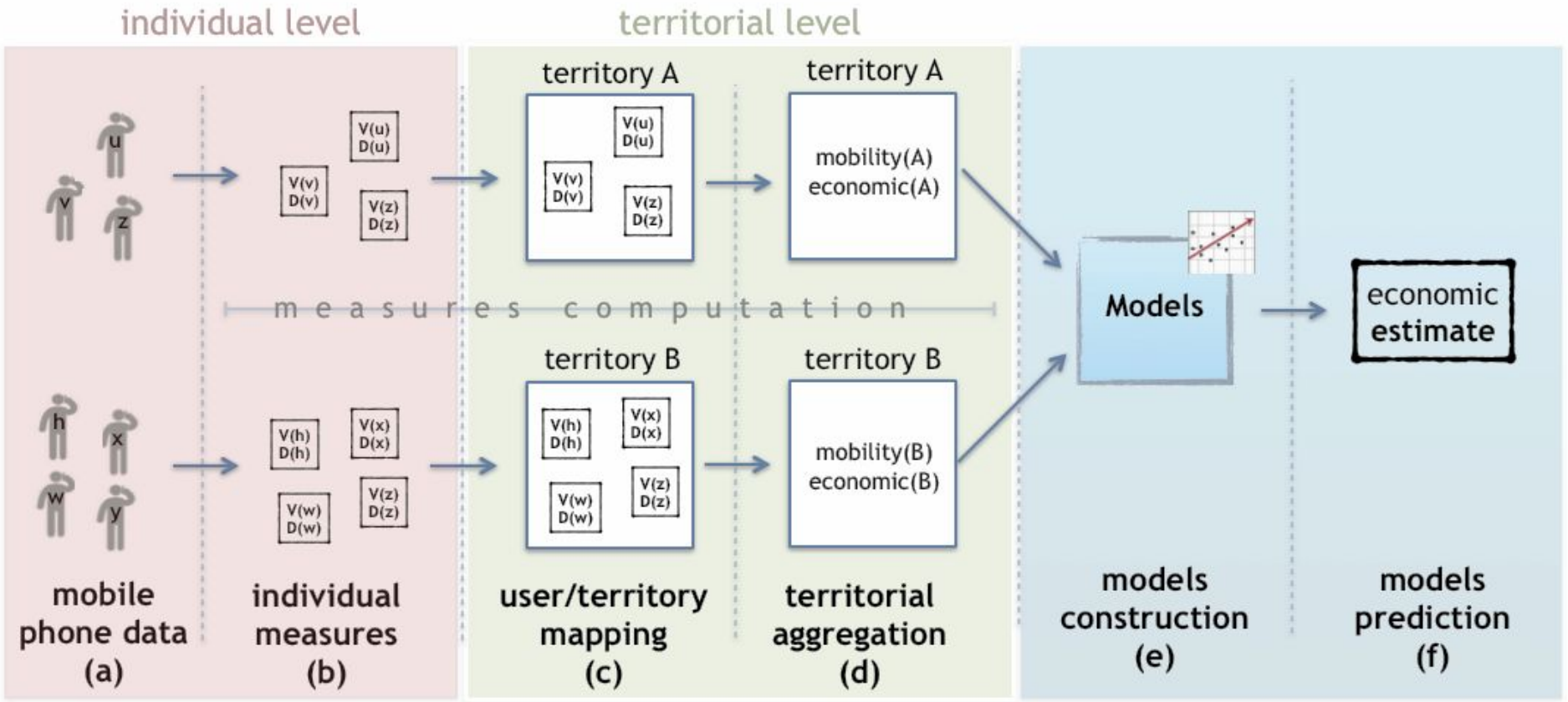


Advantages of Big Data

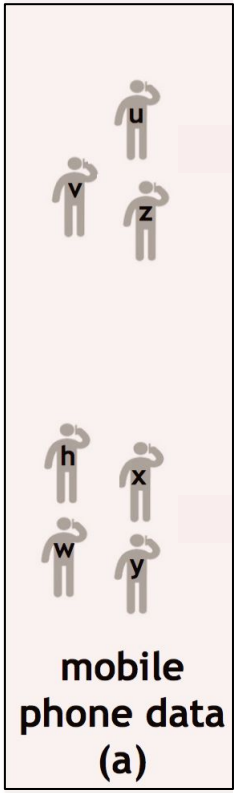
- They capture the complexity of social systems in their entirety

- They allow for the observation of complex phenomena, like diversity, resilience and equality

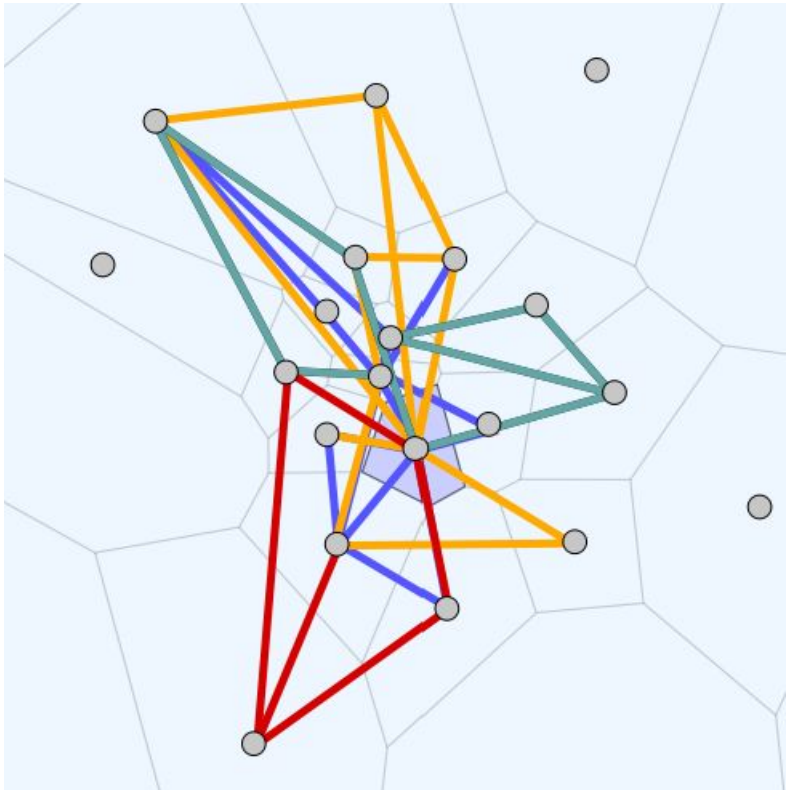
An analytical framework



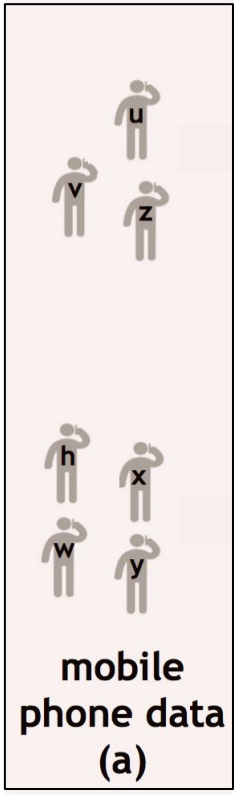
Data format



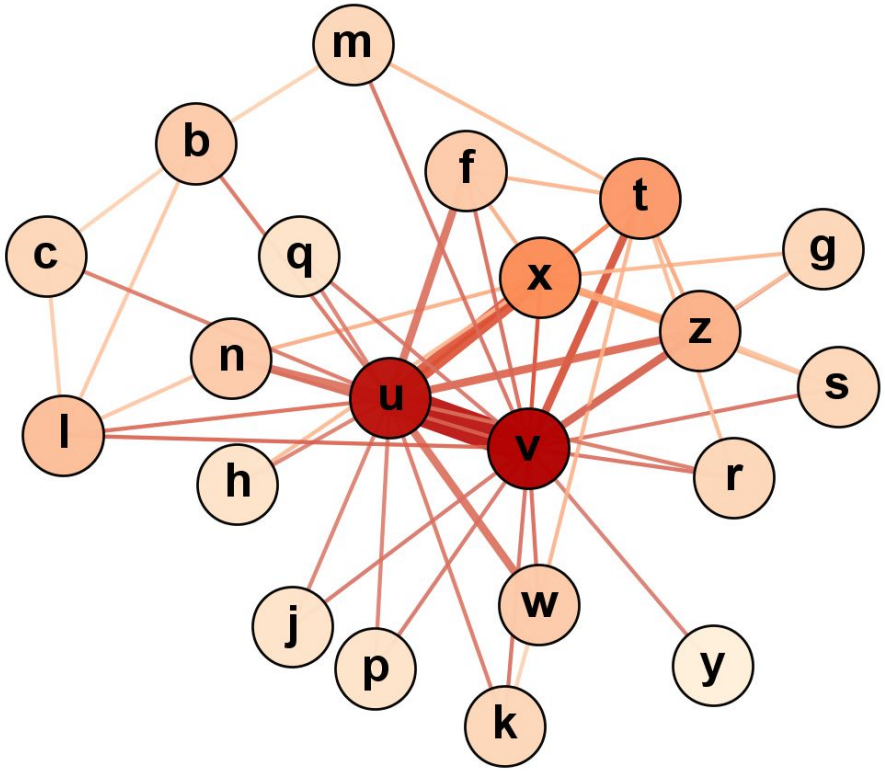
timestamp	coords	caller	callee
04/01 23:45:00	132.56, 23.64	145323	452300
04/02 06:02:00	143.28, 54.22	145323	5602
…	…	…	…

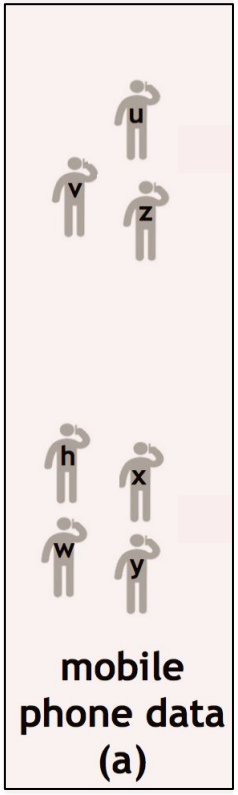


Data format



timestamp	coords	caller	callee
04/01 23:45:00	132.56, 23.64	145323	452300
04/02 06:02:00	143.28, 54.22	145323	5602
…	…	…	…





Orange dataset

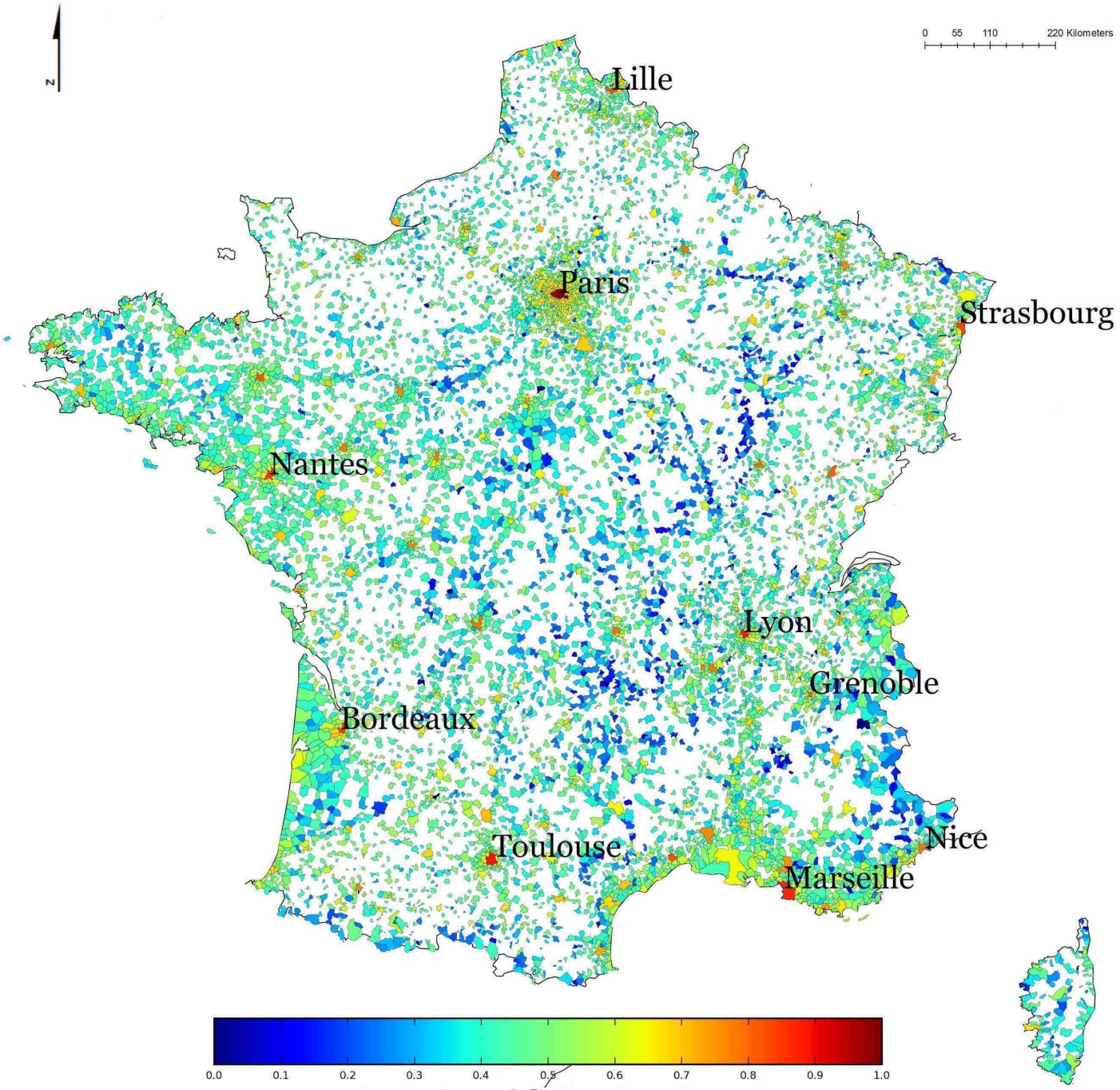
20M users

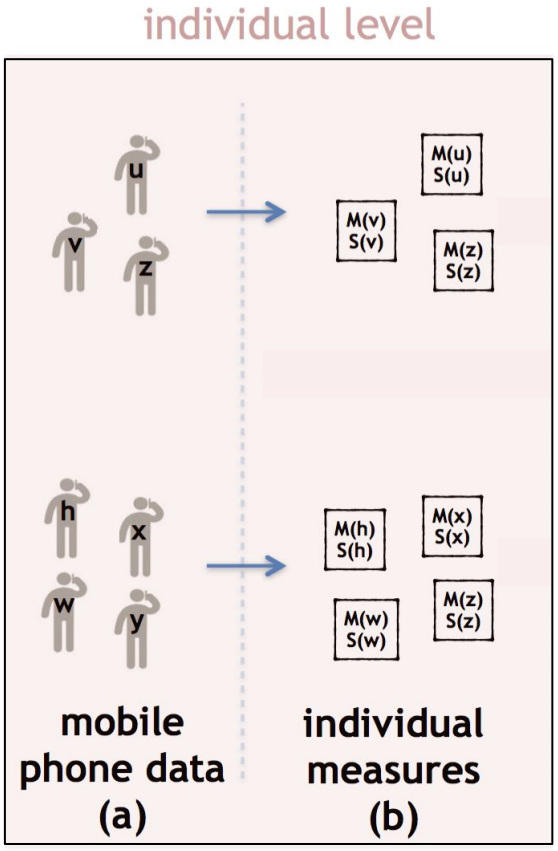
5.7G calls

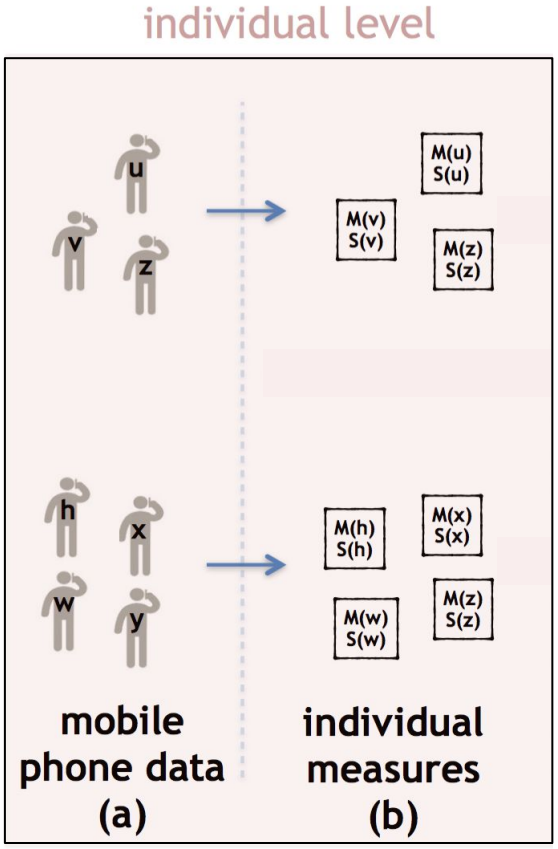
87K towers

45days

orange™


N
0 55 110 220 Kilometers
Lille
Paris
Strasbourg
Nantes
Lyon
Grenoble
Bordeaux
Toulouse
Nice
Marseille
0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0
number of Orange users

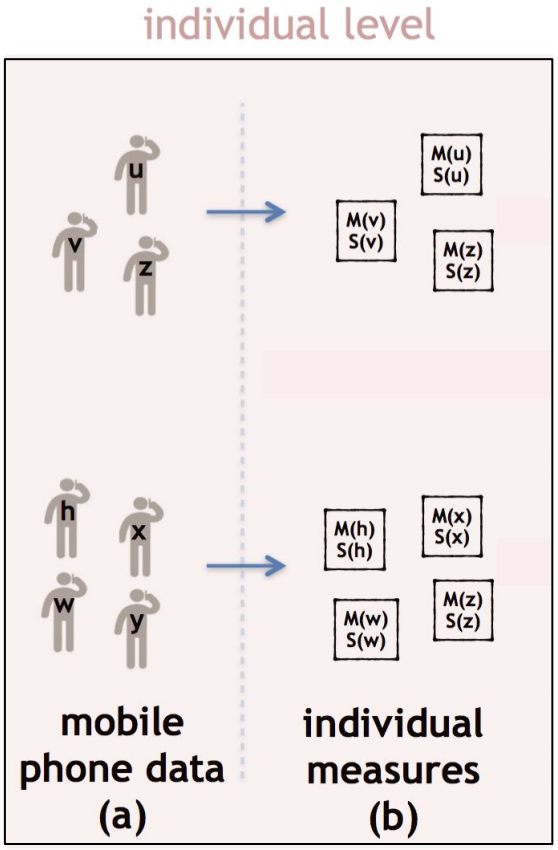

individual level
u
v
z
M(u)
S(u)
M(v)
S(v)
M(z)
S(z)
h
x
w
y
M(h)
S(h)
M(x)
S(x)
M(w)
S(w)
M(z)
S(z)
mobile phone data (a)
individual measures (b)



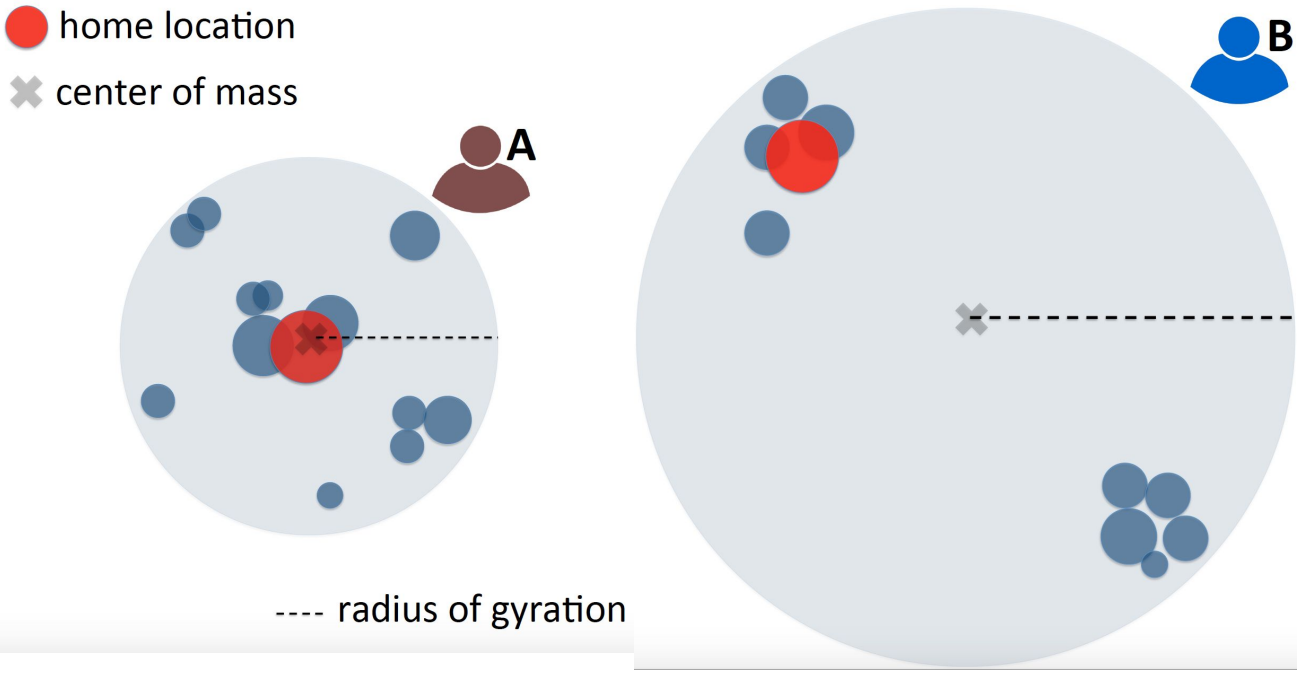
– Mobility volume:
characteristic distance traveled by an individual

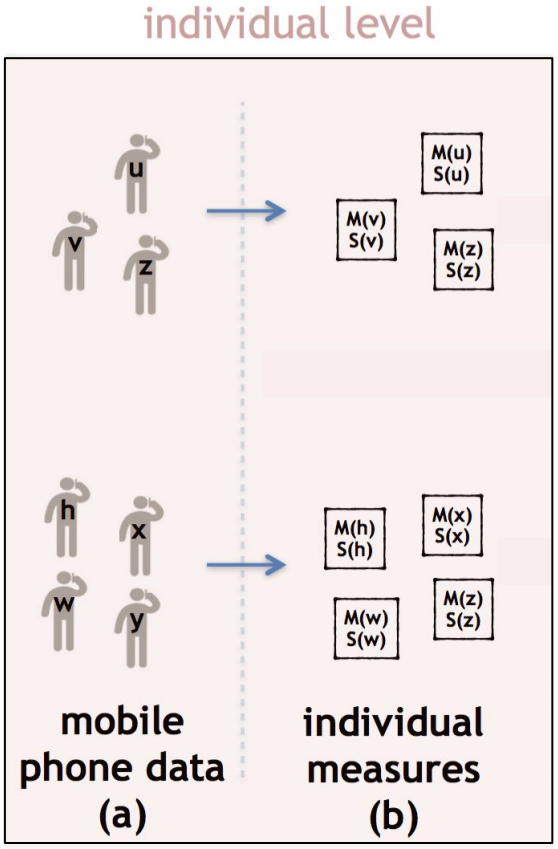
$$r_g = \sqrt{\frac{1}{N}\sum_{i=1}^{N} w_i (r_i - r_{cm})^2}$$

$$r_{cm} = \frac{1}{N}\sum_{i=1}^{n} w_i r_i \qquad N = \sum_{i=1}^{n} w_i$$



– Mobility volume:
characteristic distance traveled by an individual



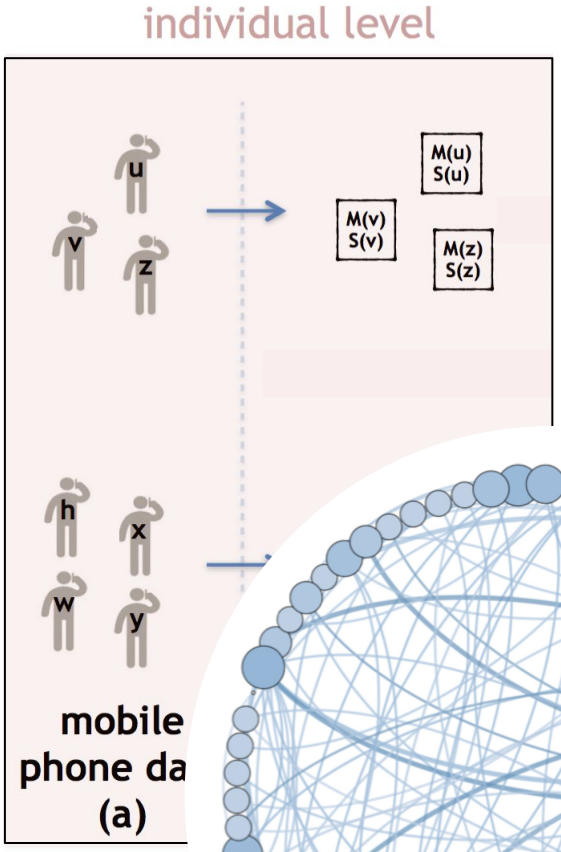


– Mobility diversity

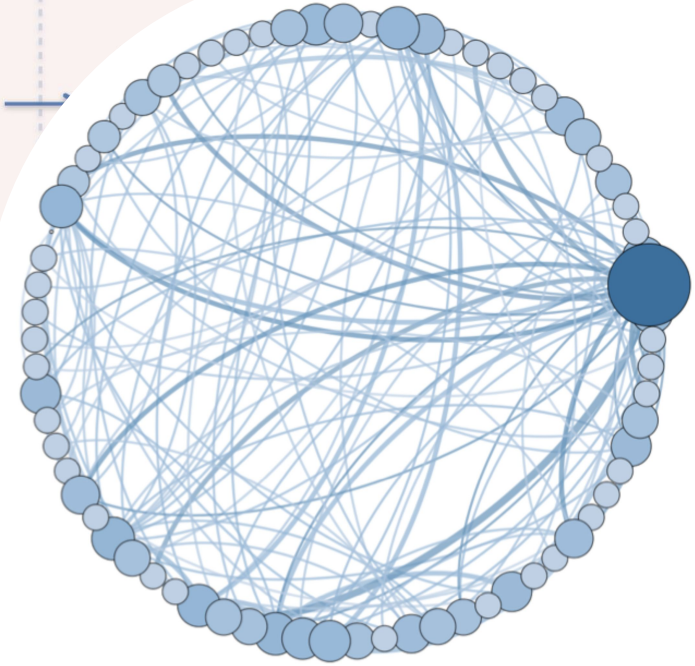
predictability of an individual's movements

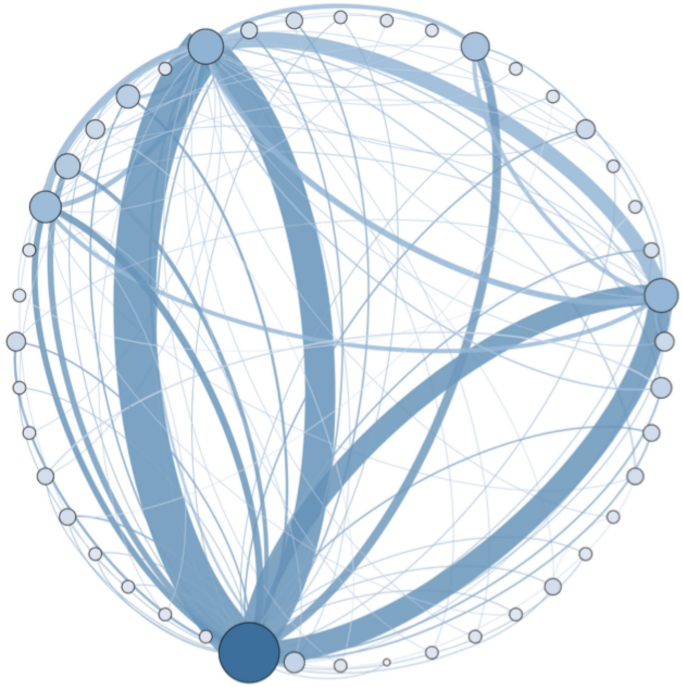
$$S^{unc} = -\sum_{i=1}^{n} p_i \log_2 p_i$$

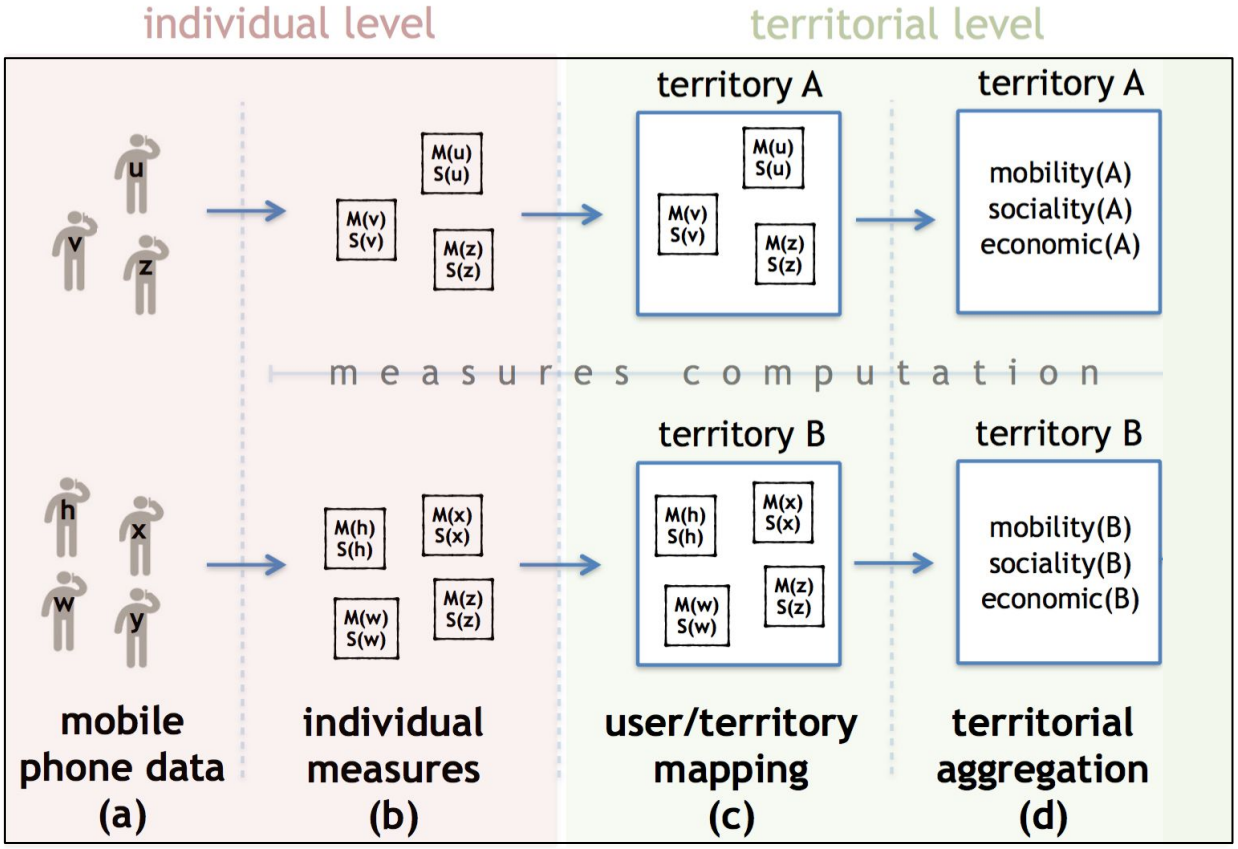
w_i/N



– Mobility diversity
predictability of an individual's movements







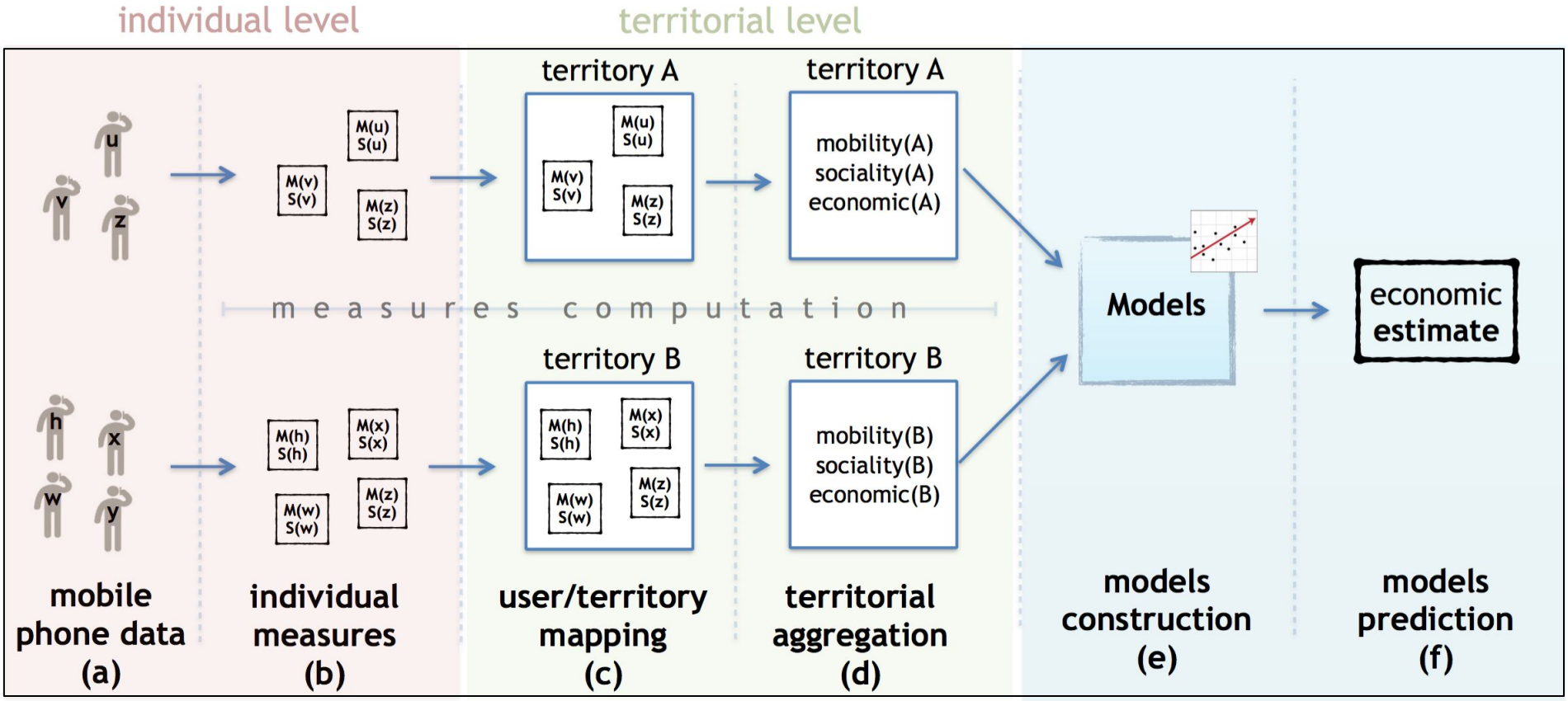
- Home location is the most frequent during nighttime (8 pm - 3 am)

- Every individual's home is assigned to its municipality

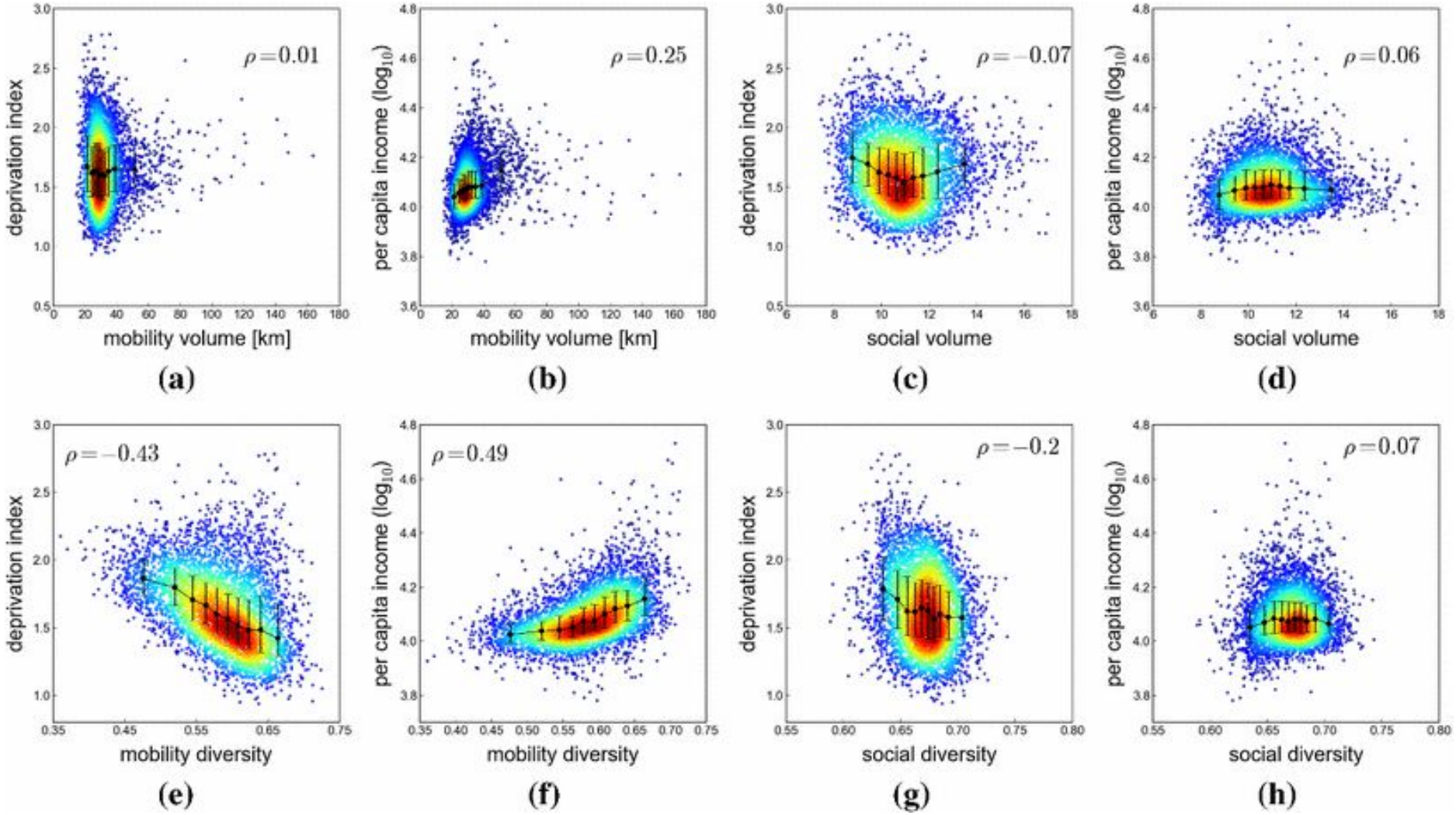
- Measures are aggregated at municipality level

- Economic measures are collected

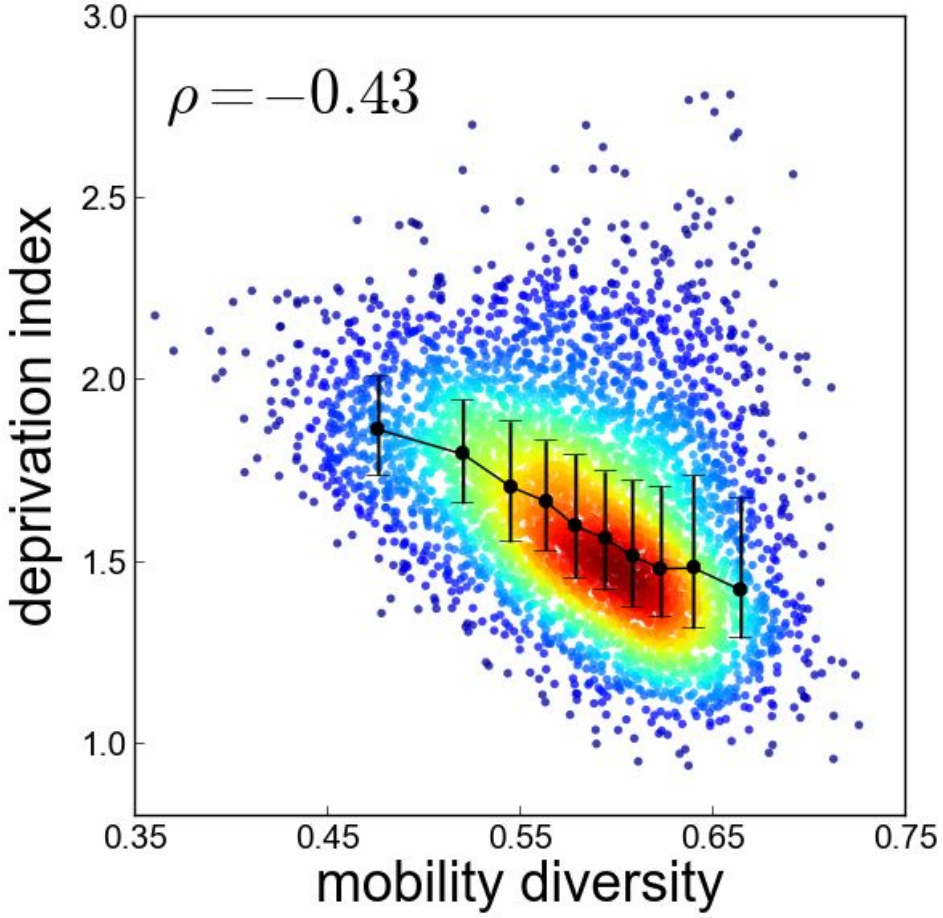
- Deprivation index
- Per capita income

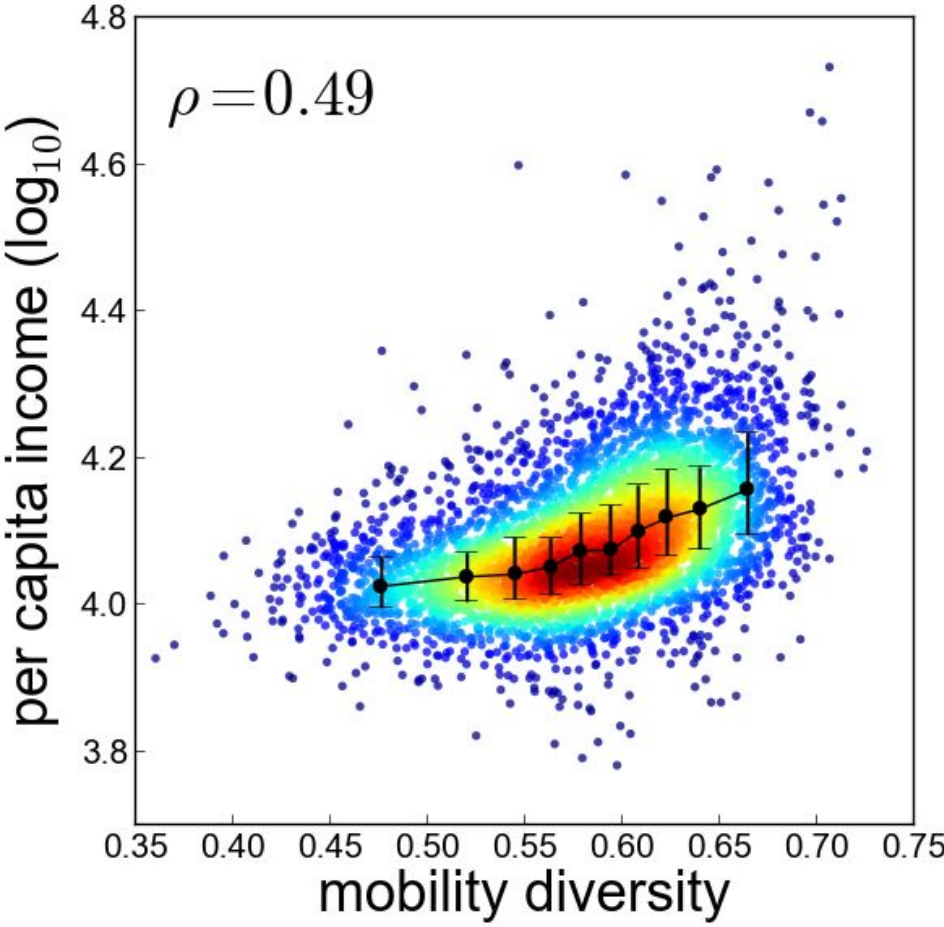

individual level
territorial level
territory A
territory A
u
v
z
M(u) S(u)
M(v) S(v)
M(z) S(z)
mobility(A)
sociality(A)
economic(A)
measures computation
Models
economic estimate
territory B
territory B
h
x
w
y
M(h) S(h)
M(x) S(x)
M(w) S(w)
M(z) S(z)
mobility(B)
sociality(B)
economic(B)
mobile phone data (a)
individual measures (b)
user/territory mapping (c)
territorial aggregation (d)
models construction (e)
models prediction (f)

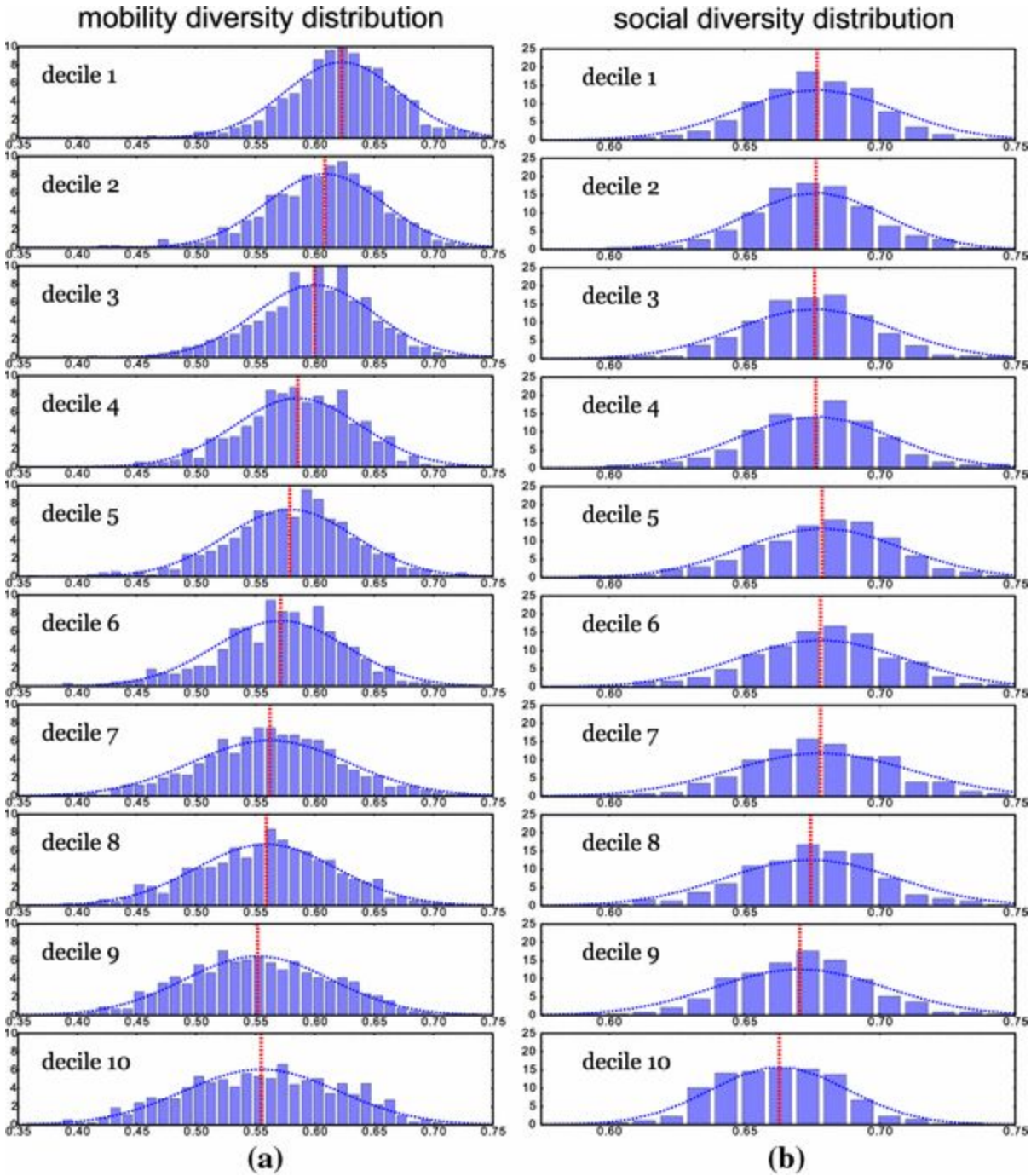
Mobility diversity and well-being



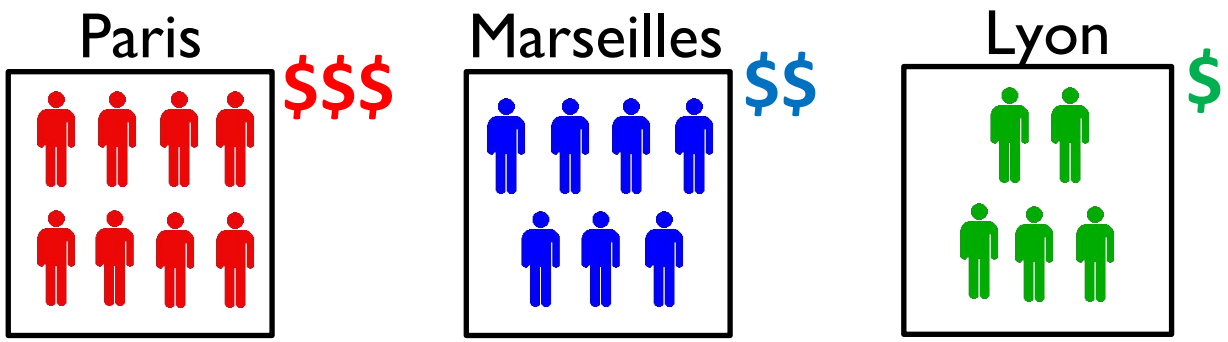
Mobility diversity and well-being



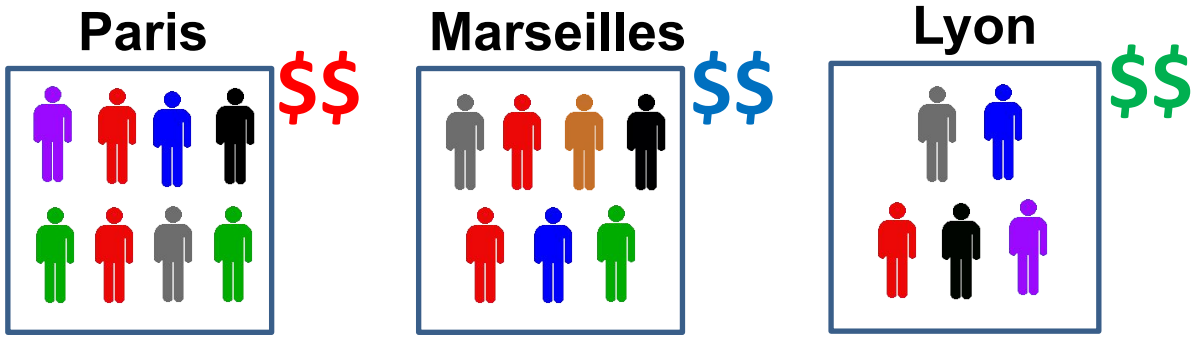



mobility diversity distribution
social diversity distribution
decile 1
decile 2
decile 3
decile 4
decile 5
decile 6
decile 7
decile 8
decile 9
decile 10
(a)
(b)

Data


Paris
$$$
Marseilles
$$
Lyon
$

NM1

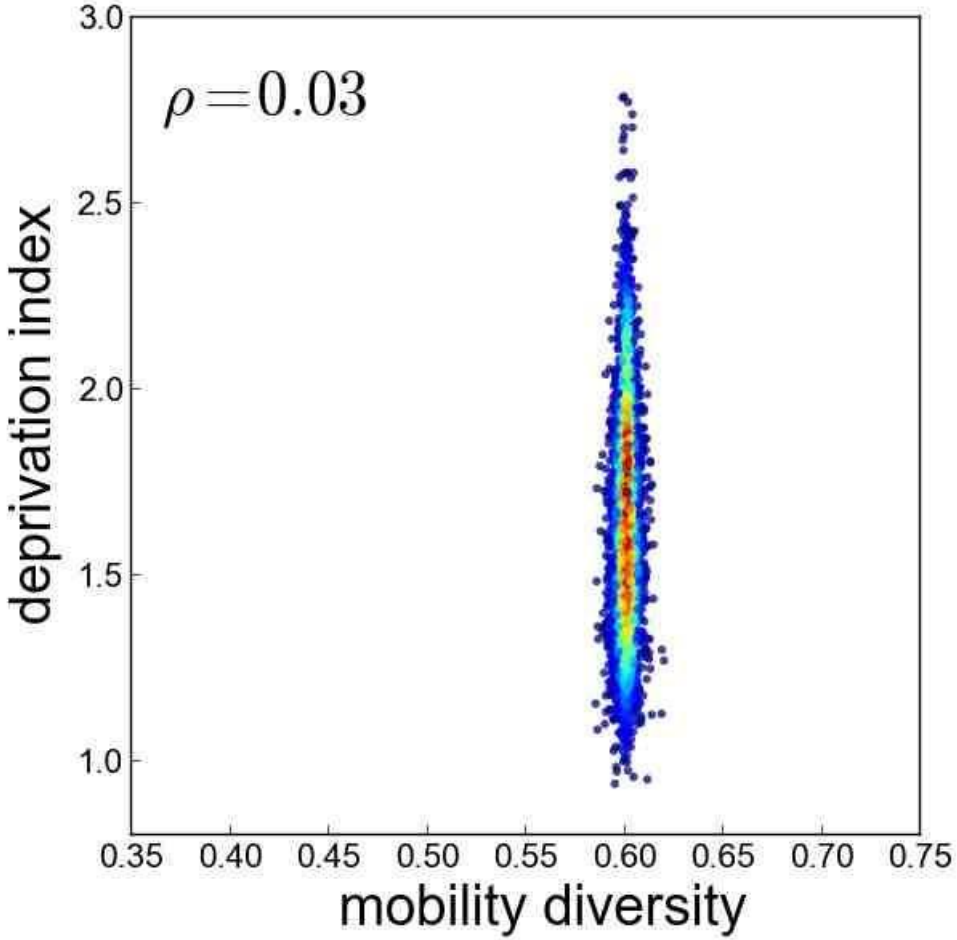

Paris
$$
Marseilles
$$
Lyon
$$

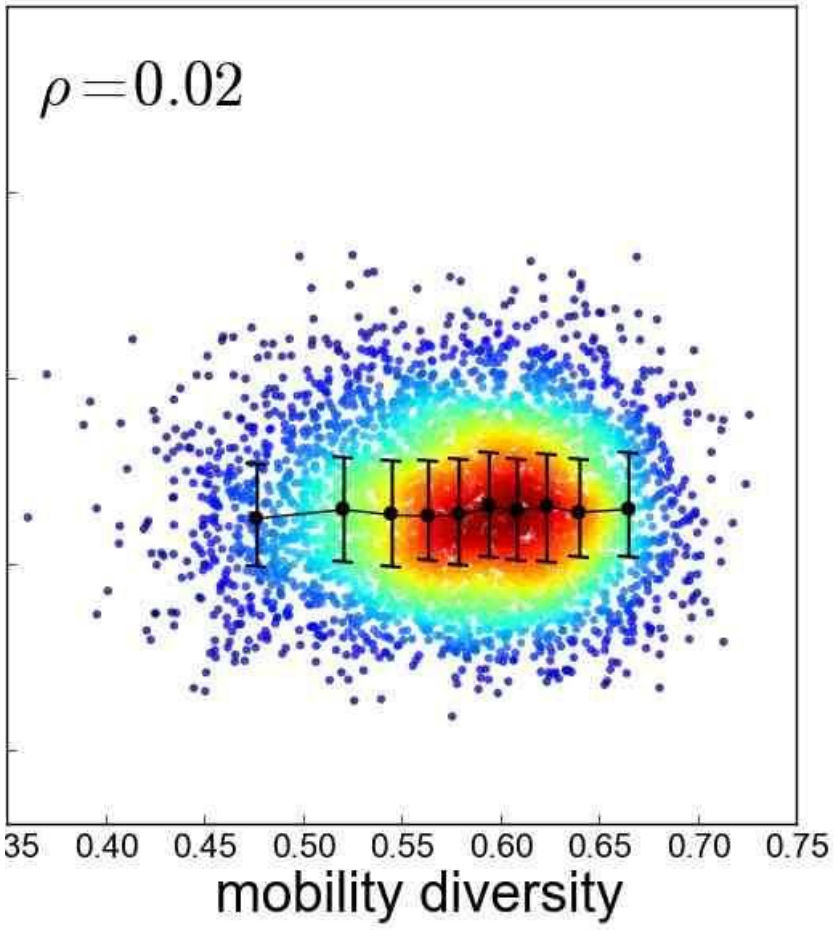
NM2


Paris
$$
Marseilles
$$
Lyon
$$

NM1

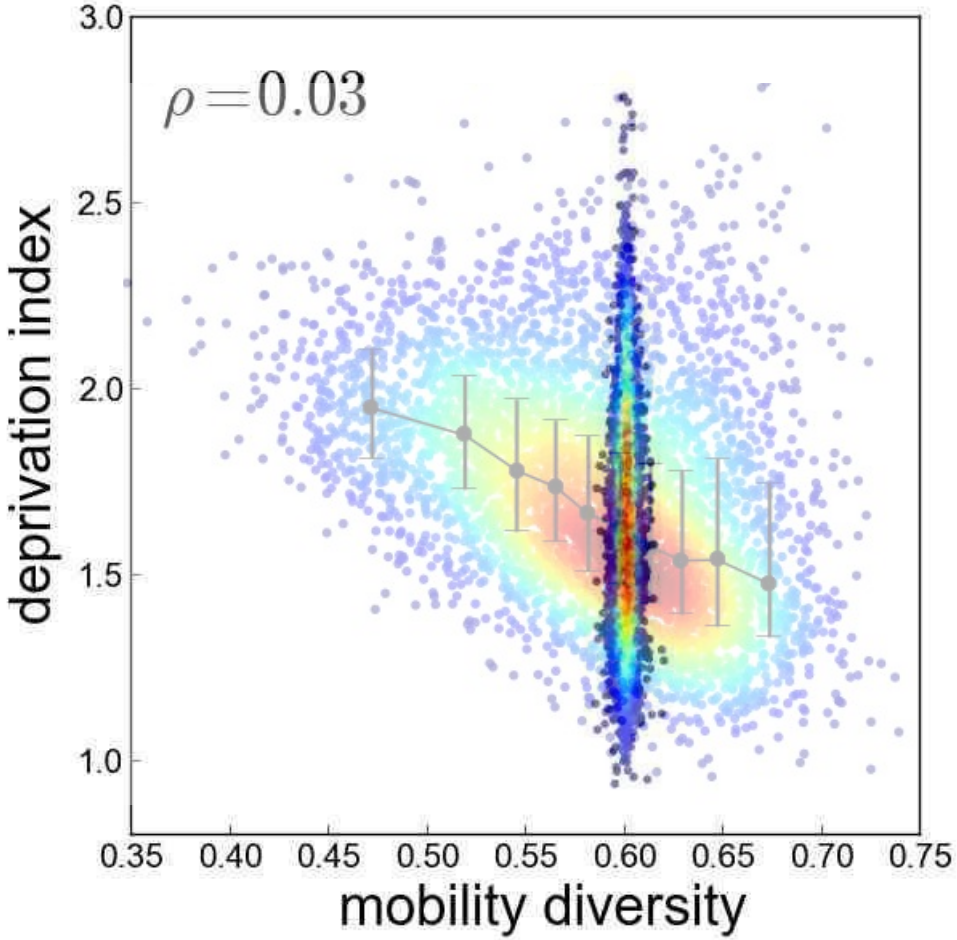
NM2

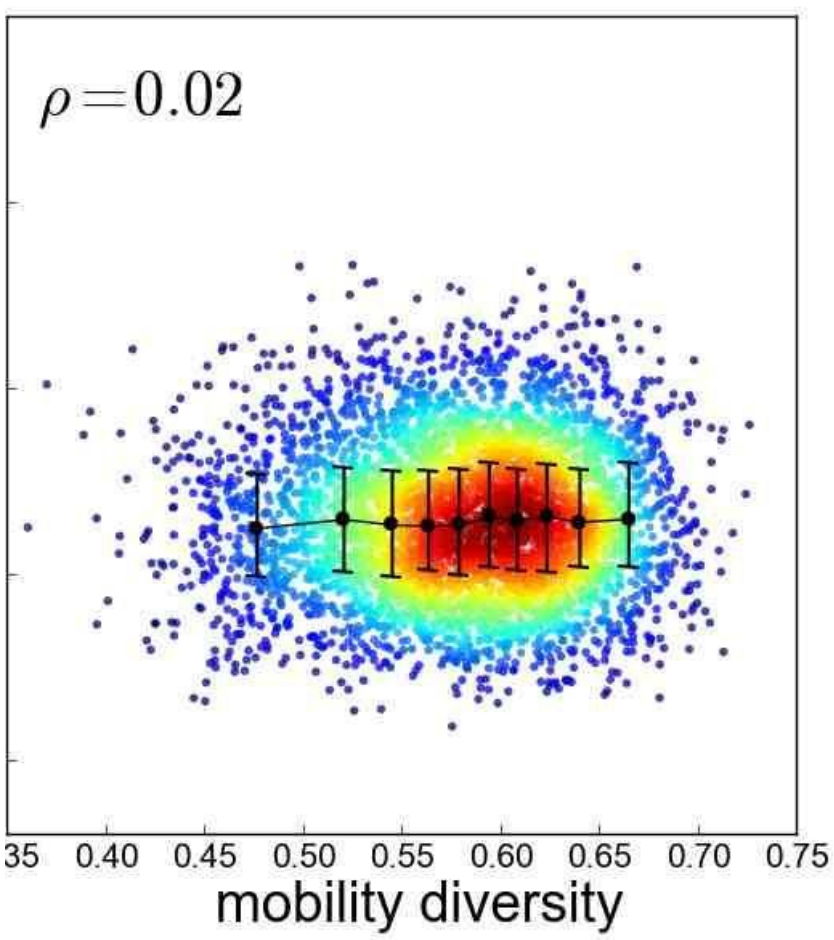

$\rho = 0.03$
deprivation index
3.0
2.5
2.0
1.5
1.0
0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75
mobility diversity

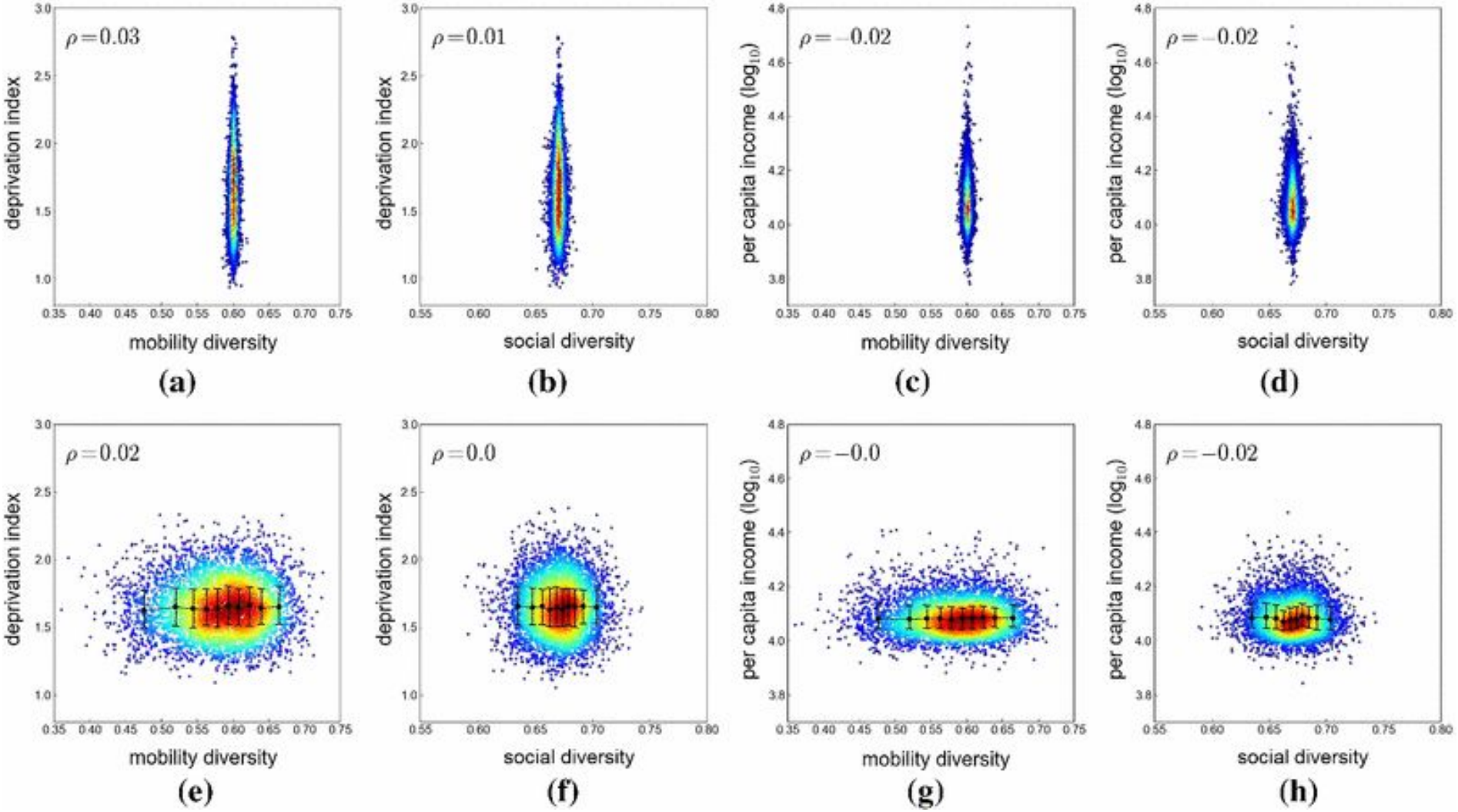

$\rho = 0.02$
35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75
mobility diversity

NM1

NM2


$\rho = 0.03$
deprivation index
3.0
2.5
2.0
1.5
1.0
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
mobility diversity


$\rho = 0.02$
35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
mobility diversity

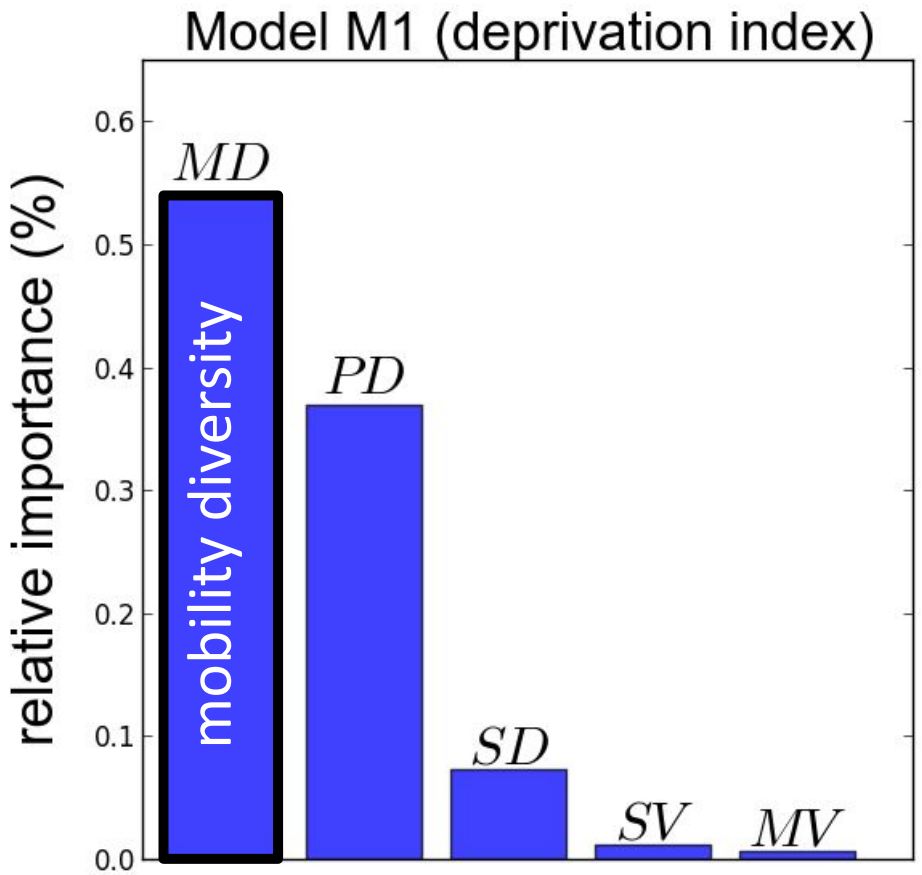

ρ = 0.03
deprivation index
mobility diversity
(a)
ρ = 0.01
deprivation index
social diversity
(b)
ρ = −0.02
per capita income (log10)
mobility diversity
(c)
ρ = −0.02
per capita income (log10)
social diversity
(d)
ρ = 0.02
deprivation index
mobility diversity
(e)
ρ = 0.0
deprivation index
social diversity
(f)
ρ = −0.0
per capita income (log10)
mobility diversity
(g)
ρ = −0.02
per capita income (log10)
social diversity
(h)

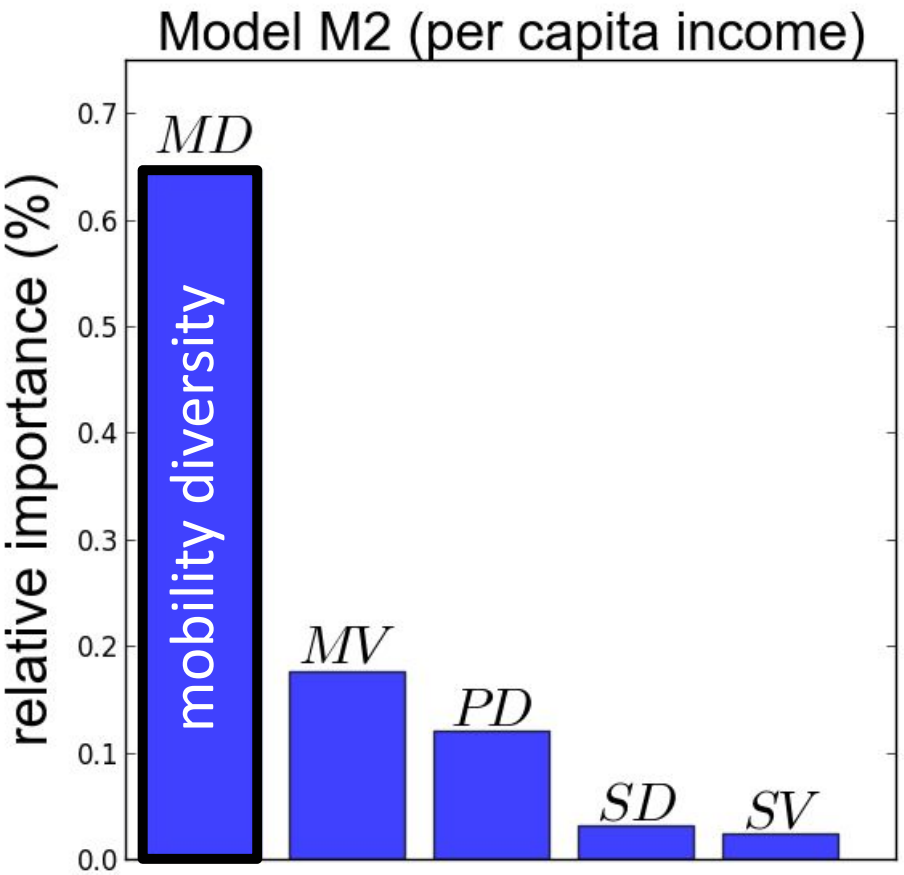
Predicting well-being

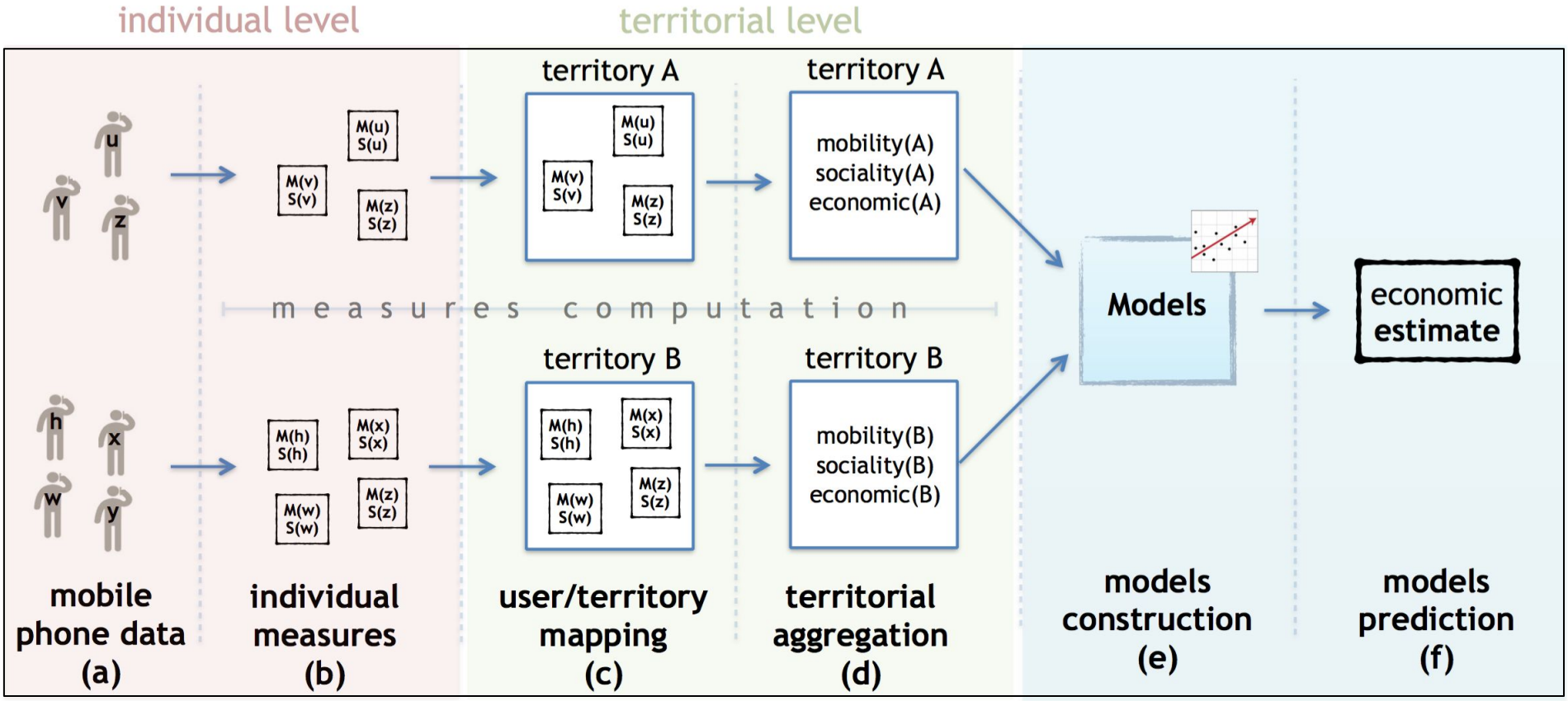
- Multivariate Regression
  predicting the exact value
  $R^2 = 0.42$ (deprivation)
  $R^2 = 0.25$ (income)

- Classification:
  predicting class of well-being (low, medium, high)
  acc = 0.61 (deprivation)
  acc = 0.54 (income)

Diversity matters

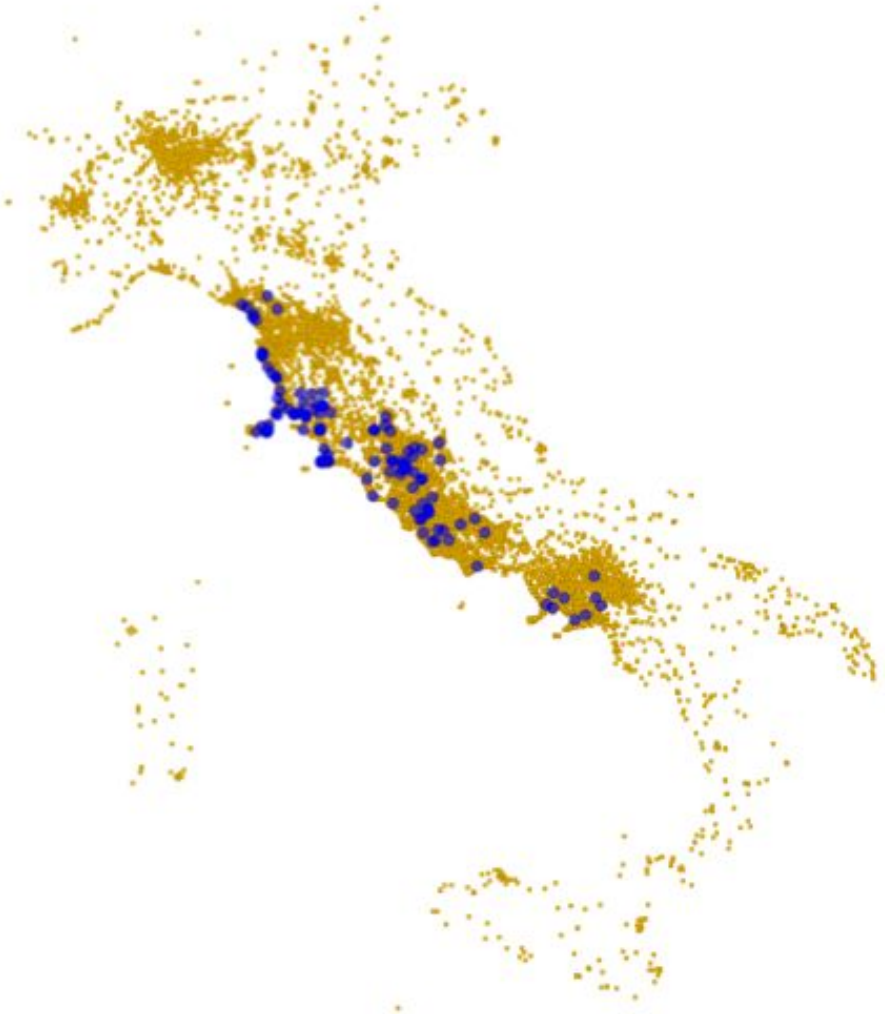




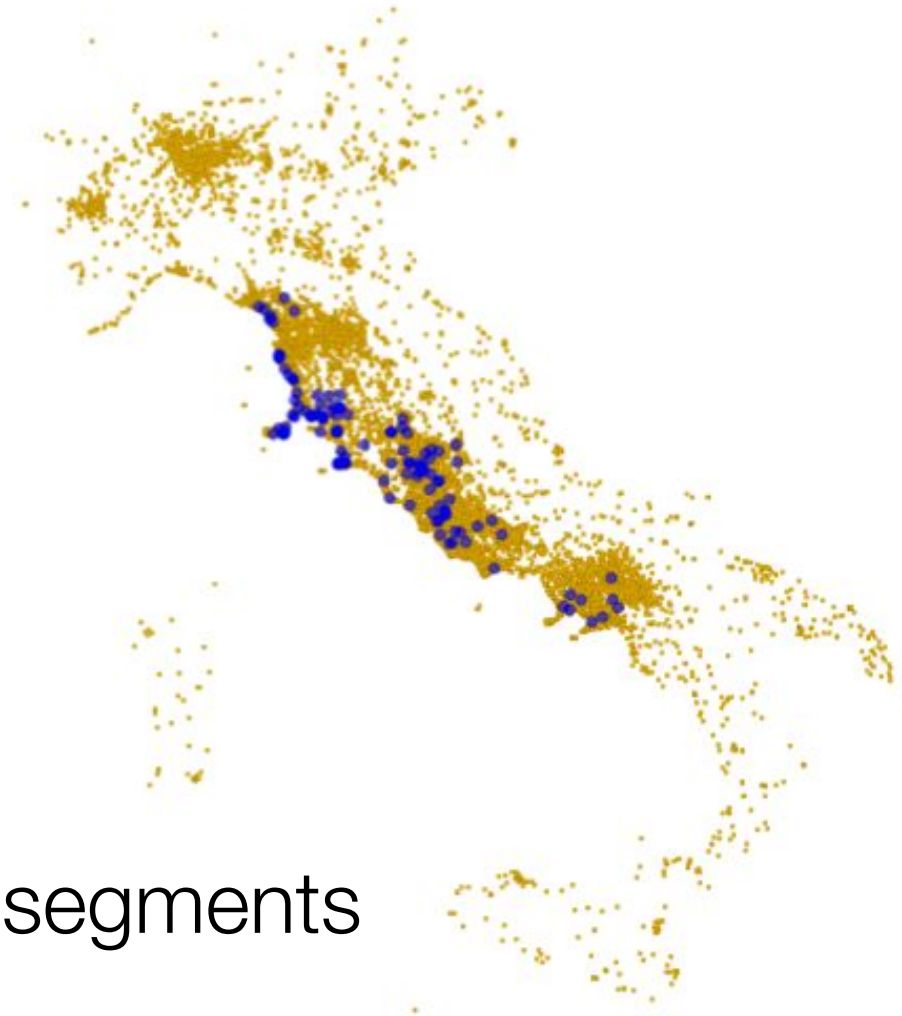

individual level
territorial level
u
v
z
h
x
w
y
M(u) S(u)
M(v) S(v)
M(z) S(z)
M(h) S(h)
M(x) S(x)
M(w) S(w)
territory A
territory B
mobility(A)
sociality(A)
economic(A)
mobility(B)
sociality(B)
economic(B)
measures computation
Models
economic estimate
mobile phone data (a)
individual measures (b)
user/territory mapping (c)
territorial aggregation (d)
models construction (e)
models prediction (f)

Wealth

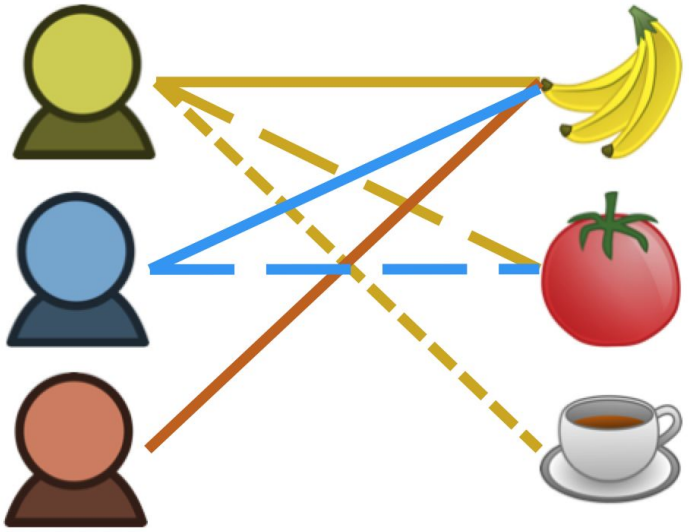
Retail market



Monitoring GDP with retail data



- 120 stores
- 1 million customers
- 300K items → 4500 segments
- from 2007 to now



We compute a measure of sophistication of customers and products

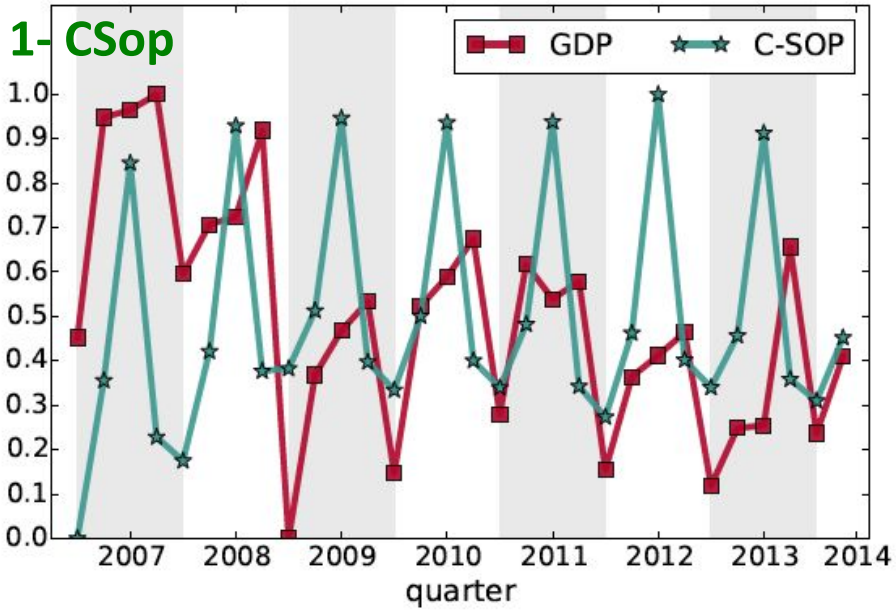
(inspired by Google PageRank)

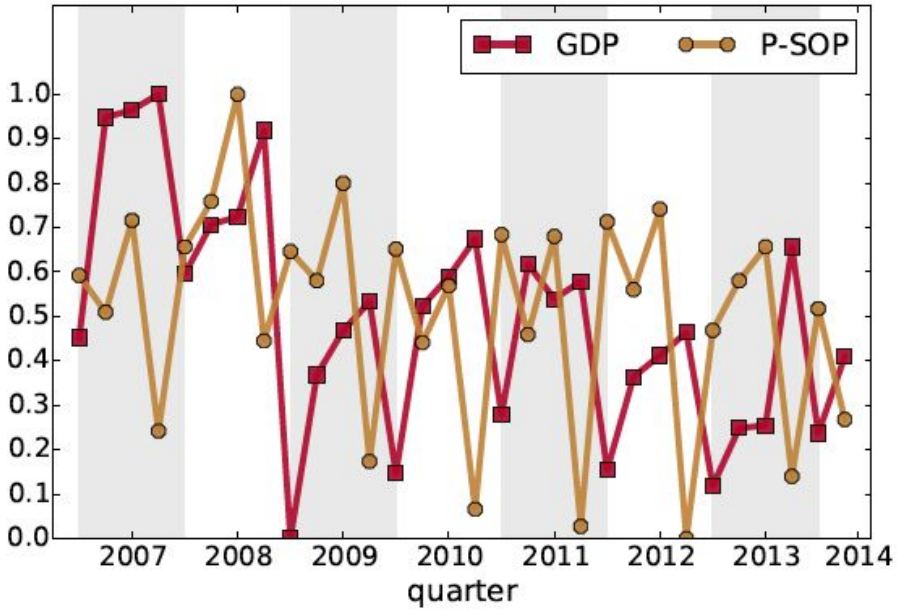
GDP and sophistication

- Rileviamo una relazione tra la sofisticatezza di clienti e prodotti e il PIL

-2 shift

1- CSop





Health

Eating habits

- 80K students
- 6 years

Meals	10,034,413
Students	82,871
with grant	*19,141*
free meals	*4,730,658*
Dishes	950
Food categories	41
Period	2,551 days
from	*01/01/2010*
to	*12/26/2016*

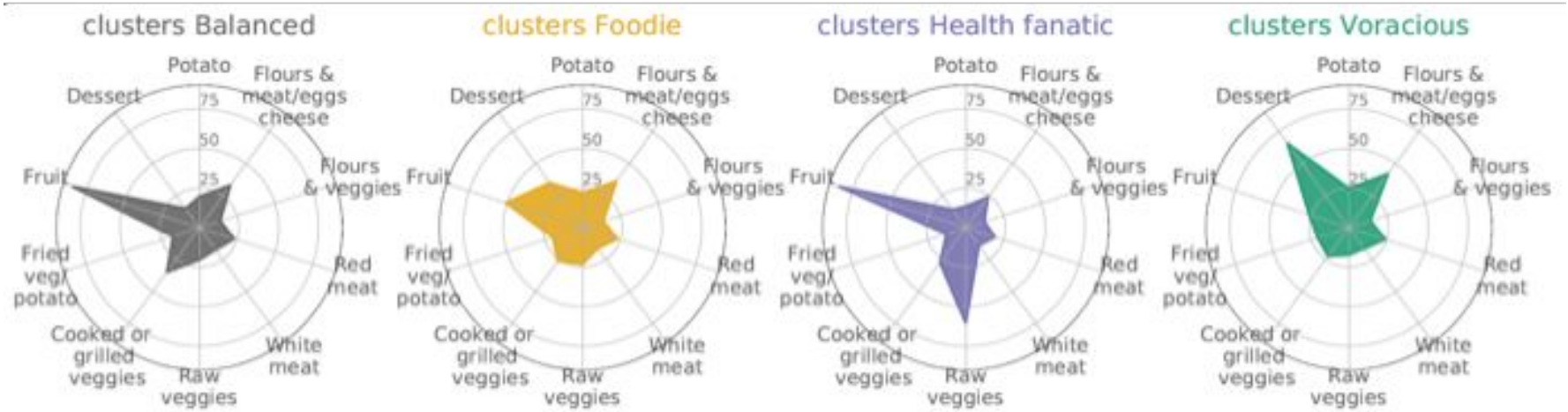
dish	attribute	description	long description
Bolognese pasta	a_{11}	meat flours	flours (pasta, couscous, dumplings) with meat/cheese/eggs
Pasta with pesto	a_{13}	veg flours	flours (pasta, couscous, dumplings) with vegetables
Pasta with zucchini	a_{13}	veg flours	flours (pasta, couscous, dumplings) with vegetables
...	...	...	...
Saffron and potato soup	a_{34}	legumes soup	potato and legumes soups
Hamburger with mushrooms	a_{51}	read meat	red meat / salami
...	...	...	...
Green salad	a_{81}	raw veg	raw vegetables
...	...	...	...
...	...	...	...
Fruit	a_{415}	fruit	fruit
Cheesecake	a_{416}	dessert	dessert

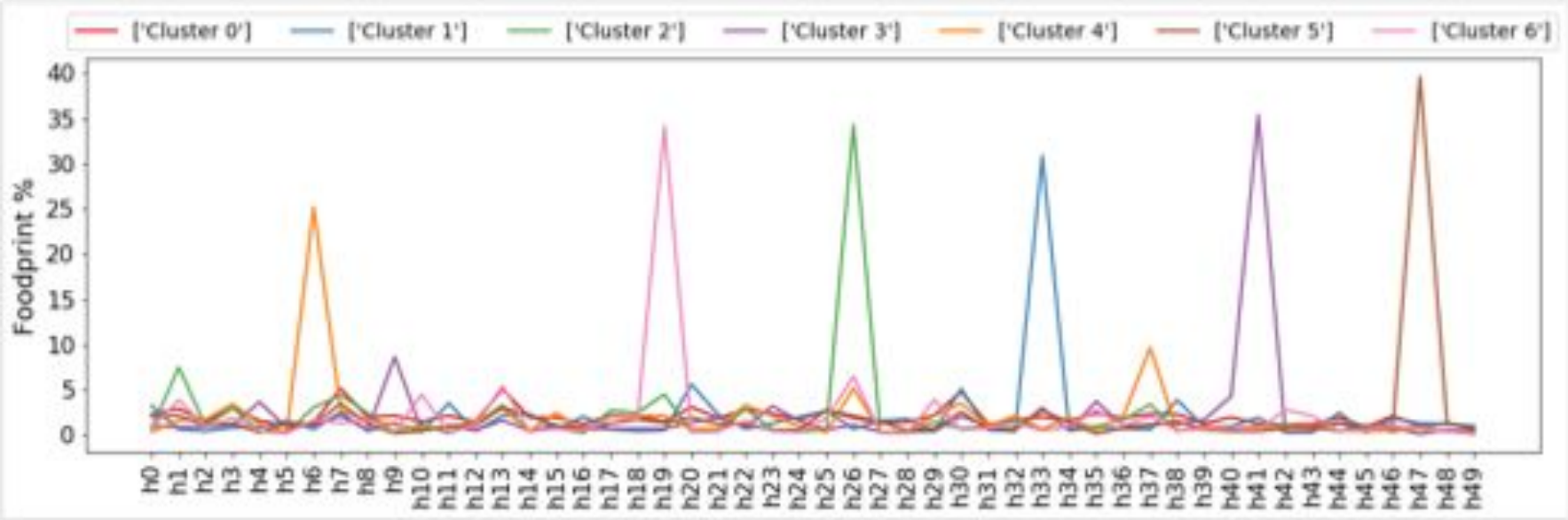
Student_id	timestamp	Dishes
A4578A	18/04/2015 12:42:00	pasta with tomato sauce, chicken breast, fruit
G23T20	18/04/2015 12:43:00	mushroom risotto, salad, fruit
GE54Y7	18/04/2015 12:44:01	pasta with tomato, fruit

Health

- Obesity is related to GERD
- Students with GERD: consumes no legumes
- Regular students with GERD: consumes -10% pasta and rice

Health


clusters Balanced
clusters Foodie
clusters Health fanatic
clusters Voracious
Potato
Flours & meat/eggs cheese
Flours & veggies
Red meat
White meat
Raw veggies
Cooked or grilled veggies
Fried veg/ potato
Fruit
Dessert
75
50
25


['Cluster 0']
['Cluster 1']
['Cluster 2']
['Cluster 3']
['Cluster 4']
['Cluster 5']
['Cluster 6']
Foodprint %
40
35
30
25
20
15
10
5
0
h0 h1 h2 h3 h4 h5 h6 h7 h8 h9 h10 h11 h12 h13 h14 h15 h16 h17 h18 h19 h20 h21 h22 h23 h24 h25 h26 h27 h28 h29 h30 h31 h32 h33 h34 h35 h36 h37 h38 h39 h40 h41 h42 h43 h44 h45 h46 h47 h48 h49

In summary ...

- Analytical frameworks can be defined to extract complex indeces from Big Data
- Associations between these indices and well-being indicators can be investigated
- Predictive models can be constructed to nowcast well-being indicators
- This overcomes the limitations of surveys and censuses