

Probabilistic record linkage of Brazil's public healthcare data: case study on a cohort of 103 million records

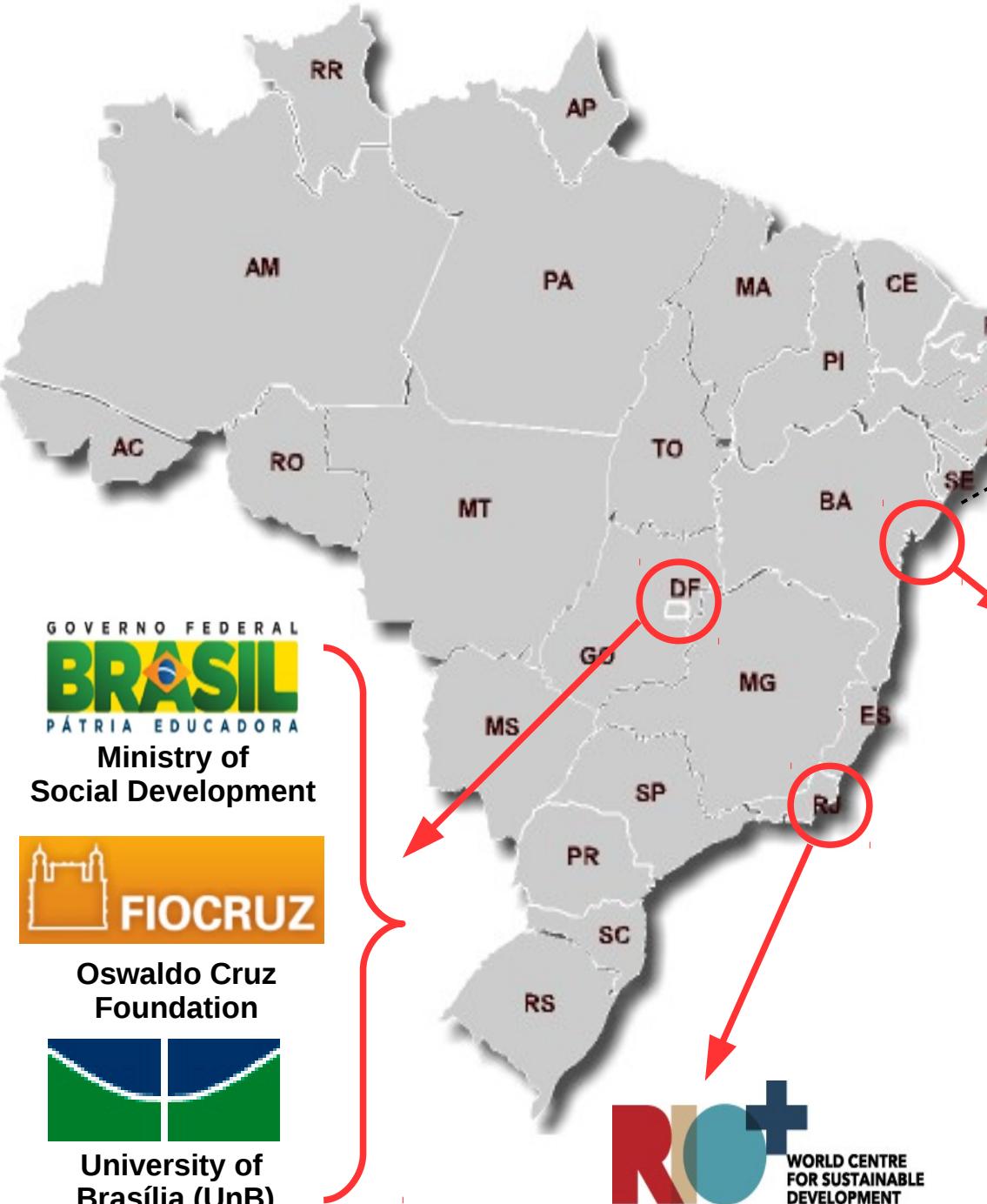
Marcos E. Barreto

Professor

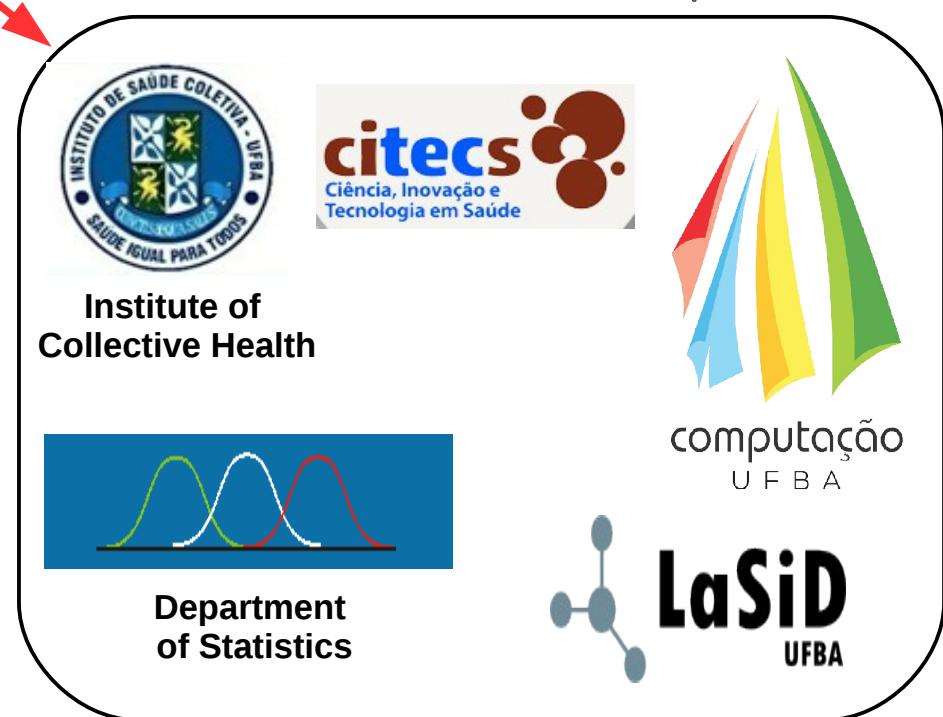
Distributed Systems Laboratory (LaSiD)
Computer Science Department (DCC)
Federal University of Bahia (UFBA)

Farr Institute and
London School of Hygiene and Tropical Medicine
London, May 2015

People involved



Federal University
of Bahia (UFBA)



Outline

- Brazilian health system and social programmes
- Project overview
- Spark-based pipeline for record linkage
- Cohort design and first steps

Outline

- Brazilian health system and social programmes
- Project overview
- Spark-based pipeline for record linkage
- Cohort design and first steps

Brazil – some key facts

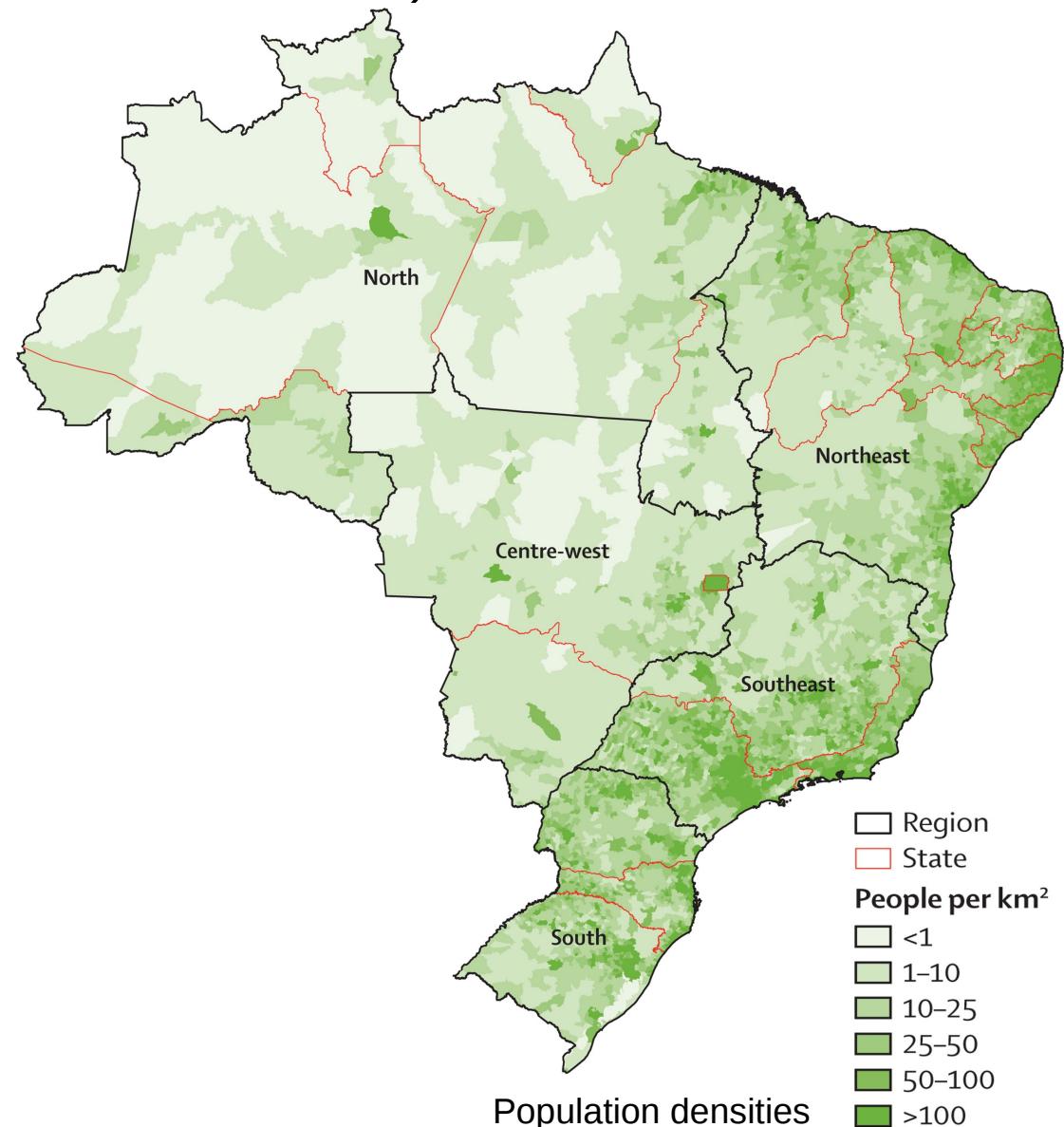
- World's fifth most populous country.
 - 190,732,694 inhabitants (IBGE census - 2010)
 - 5,563 municipalities
- Very irregular population densities.
- Widespread regional and social inequalities.

Northeast and North

- ✗ 1st and 2nd poorest regions
- ✗ lowest population density

Southeast region

- ✗ 11% of territory
- ✗ 43% of population
- ✗ 56% of gross domestic product



The Brazilian health system: history, advances, and challenges.

Jamilson Paim, Cláudia Travassos, Celia Almeida, Ligia Bahia, James Macinko

The Lancet - Volume 377, Issue 9779, Pages 1778-1797 (May 2011)

DOI: 10.1016/S0140-6736(11)60054-8

=> historical development and components of the Brazilian health system, focusing on the period 1970-2010.

1988 - 2010

Health System	Key health challenges
<ul style="list-style-type: none">• Creation of the SUS• Decentralisation of the health system <p>9th National Health Conference</p> <ul style="list-style-type: none">• INAMPS repealed (1993)• Family Health Programme set up (1994)• Crisis in funding and creation of Provisional Contribution on Financial Transactions (1996)• Free treatment for HIV/AIDS through the SUS• Per head PHC funding (1998)• 10th and 11th National Health Conferences• Health-care operating norms and regionalisation established• Regulation of the private health plans• National Health Surveillance Agency set up (1999)• Supplementary Health Care Agency set up to regulate and oversee private health plans (2000)• The generic drugs law passed• The Arouca Law instituted indigenous health care as part of the SUS• Constitutional amendment addressed the instability in SUS financing and defined the duties of the Union, states, and municipalities (2000)• Psychiatric reform law passed (2001)• Expansion and consolidation of PHC• Mobile emergency care (ambulance) system set up (2003)• Pact for Health established (Pact in Defence of the SUS, Management Pact, the Pact for Life; 2006)• National Primary Care policy (2006)• Health Promotion (2006)• 12th and 13th National Health Conferences• National Commission on Social Determinants of Health and National Oral Health Policy (Brasil Soridente; 2006)• 24-h emergency care units set up in municipalities with populations >100 000 (2008)• Multi-professional Family Health Support Teams set up to support the Family Health Programme (2008)	<p>Cholera and dengue fever epidemics, mortality from external causes (mostly homicides and traffic accidents) Cardiovascular disease most common cause of death, followed by external causes and cancers</p> <p>Decrease in infant mortality, no change in prevalence of tuberculosis, stabilisation in prevalence of AIDS-rates illness, increase in prevalence of dengue fever, and increase in incidence of visceral leishmaniasis and malaria</p> <p>Life expectancy was about 72·8 years (68·7 for men and 76·4 for women) at the start of the 21st century</p> <ul style="list-style-type: none">• Infant mortality rate was 20·7 per 1000 livebirths (2006)• Decrease in the prevalence of Hansen's disease and immunisation-preventable diseases• Life expectancy increased to 72·8 years (69·6 for men and 76·7 for women; 2008)

Brazilian health system

- Made up of a complex network of complementary and competitive service providers and purchasers.
- 3 subsectors:
 - Public
 - Private (for-profit and non-profit)
 - Private health insurance
- Public and private subsectors are distinct but interconnected.
- People can use services in all three subsectors, depending on ease of access and their ability to pay.

Brazilian health system

- Private subsystem (for- and non-profit)
 - Medical practices, specialist diagnostic, therapeutic clinics, private hospitals, and private health insurance companies.
 - Some services contracted-out by the public sector.
 - Responsible for most secondary and tertiary health care services.

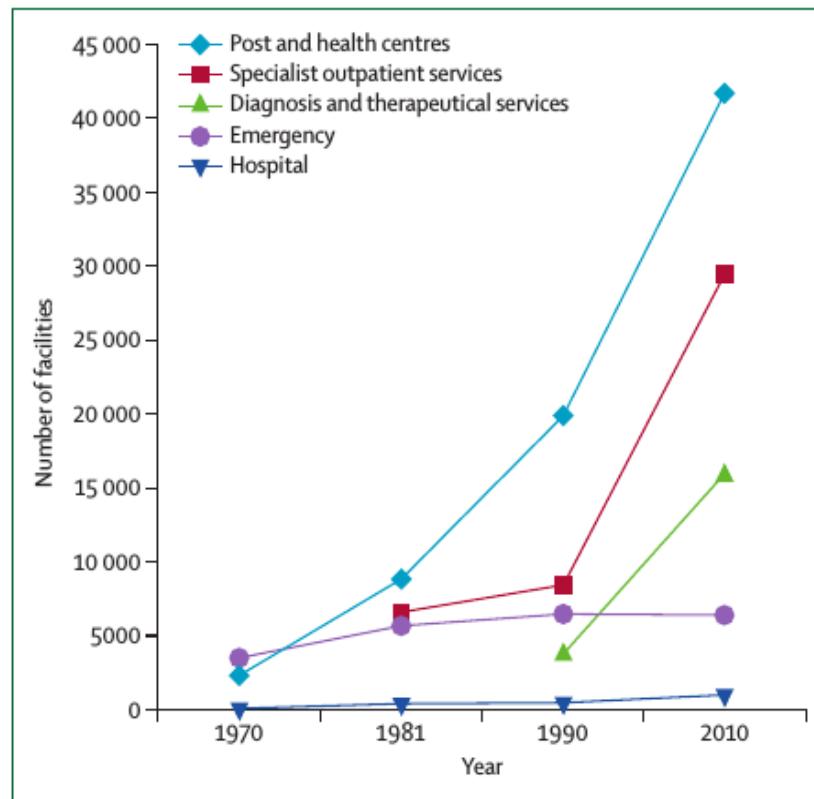


Figure 7: Type of health-care facilities in Brazil, 1970–2010

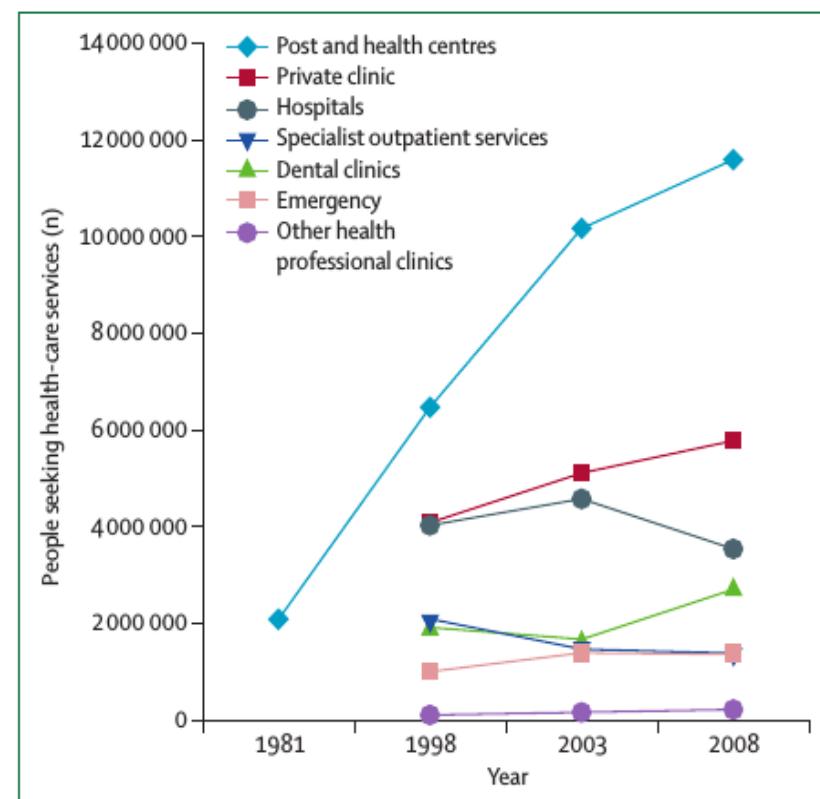


Figure 8: Health service demand by service type

Brazilian health system



- **Private subsystem (health insurance)**
 - Different healthcare providers and levels of choice.
 - Users have better access to preventive services...
 - ...but they access high cost services, complex procedures (haemodialysis and transplants), and some vaccines in the public subsystem.

Year	Private medical aid plan beneficiaries with or without dental care	Private dental-only plan beneficiaries
Dec/2003	32.074.667	4.325.568
Dec/2004	33.840.716	5.312.915
Dec/2005	35.441.349	6.204.404
Dec/2006	37.248.388	7.349.643
Dec/2007	39.316.313	9.164.386
Dec/2008	41.247.802	10.711.471
Dec/2009	42.421.531	12.839.738
Dec/2010	45.327.432	14.470.793
Dec/2011	46.974.170	16.919.583
Dec/2012	47.943.091	18.606.149



Brazilian health system

- Public subsystem
 - **Unified Health System (SUS)** implemented from 1990 onwards and tasked with:
 - undertaking health promotion / surveillance / education, and vector control.
 - ensuring continuity of care at primary, hospital, and specialist outpatient levels.
 - Funding from tax revenues and social contributions from federal, state, and municipal budgets.
 - Wide decentralisation.
 - Legislation, health councils, inter-managerial committees.
 - Important initiatives.
 - HIV/AIDS control programme, tobacco control efforts, Indigenous health programme, Mobile Emergency Care Service (SAMU), National Oral Health Policy (*Brasil Soridente*), National Supplementary Health Agency.

Brazilian health system

- Public subsystem

- **Family Health Programme (FHP)** implemented in 1994.

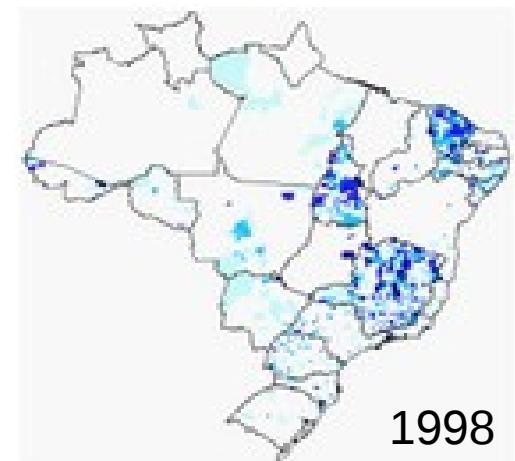
- Free community access to primary care.

- Coverage (2013):
 - 96% of municipalities, 56,4% of population.

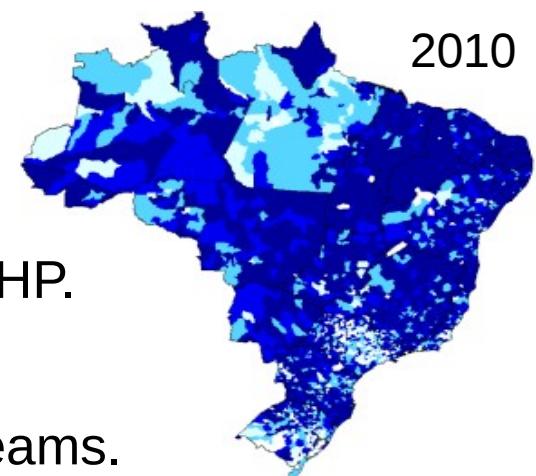
- FHP teams:
 - Physician (1), nurse (1), nursing staff (2), community health workers (CHW) (6), oral health professionals (1).
 - 1 team – 1,000 families; 1 CHW – 150 families

- Centres for Support for Family Health (NASF)
 - Created in 2008 to support the consolidation of the FHP.
 - Different arrangements (NASF1, NASF2, NASF3):
 - Multi-professional support teams linked to FHP teams.

Municipal coverage



1998



2010



Impact of the Family Health Program on infant mortality in Brazilian municipalities

Rosana Aquino, Nelson F. de Oliveira, and Maurício L. Barreto

American Journal of Public Health, 2009; 9(1):87-93

Variables	Infant Mortality Rate		Neonatal Mortality	Postneonatal Mortality
	Crude RR (95% CI)	Adjusted RR (95% CI)	Rate, RR (95% CI)	Rate, RR (95% CI)
FHP coverage				
No FHP ^a (Ref)	1.00	1.00	1.00	1.00
Incipient FHP ^b	0.84 (0.82, 0.85)	0.87 (0.86, 0.89)	0.90 (0.89, 0.92)	0.82 (0.80, 0.84)
Intermediate FHP ^c	0.77 (0.75, 0.79)	0.84 (0.82, 0.86)	0.86 (0.84, 0.89)	0.78 (0.75, 0.81)
Consolidate FHP ^d	0.68 (0.64, 0.73)	0.78 (0.73, 0.83)	0.81 (0.76, 0.88)	0.69 (0.62, 0.76)

Reducing childhood mortality from diarrhea and lower respiratory tract infection in Brazil

Davide Rasella, Rosana Aquino, and Maurício L. Barreto

Pediatrics. 2010;126:e534-40

Variables	Mortality From Diarrheal Diseases, RR (95% CI)	Mortality From Lower Respiratory Infections, RR (95% CI)	Mortality From Injuries, RR (95% CI)
Crude			
FHP coverage	1.00	1.00	1.00
No FHP ^a	0.84 (0.75–0.94)	0.84 (0.78–0.91)	1.04 (0.96–1.13)
Low ^b	0.75 (0.66–0.85)	0.75 (0.68–0.82)	0.91 (0.83–1.00)
Intermediate ^c	0.61 (0.53–0.70)	0.74 (0.66–0.83)	0.99 (0.89–1.11)
High ^d			
Adjusted^e			
FHP coverage	1.00	1.00	1.00
No FHP ^a	0.89 (0.79–1.00)	0.87 (0.80–0.94)	1.05 (0.97–1.13)
Low ^b	0.82 (0.73–0.94)	0.80 (0.72–0.88)	0.92 (0.84–1.01)
Intermediate ^c	0.69 (0.60–0.80)	0.81 (0.72–0.92)	1.01 (0.89–1.14)
No. of observations	8150	10 074	10 998
No. of municipalities	1355	1679	1833

Social programmes

- **Cadastro Único (CadÚnico)**

- Created in 2007 to keep socioeconomic data from **low-income families**.
 - Monthly income of up to half a minimum wage (\approx US\$ 131) per person.
 - Total monthly income of three minimum wages.



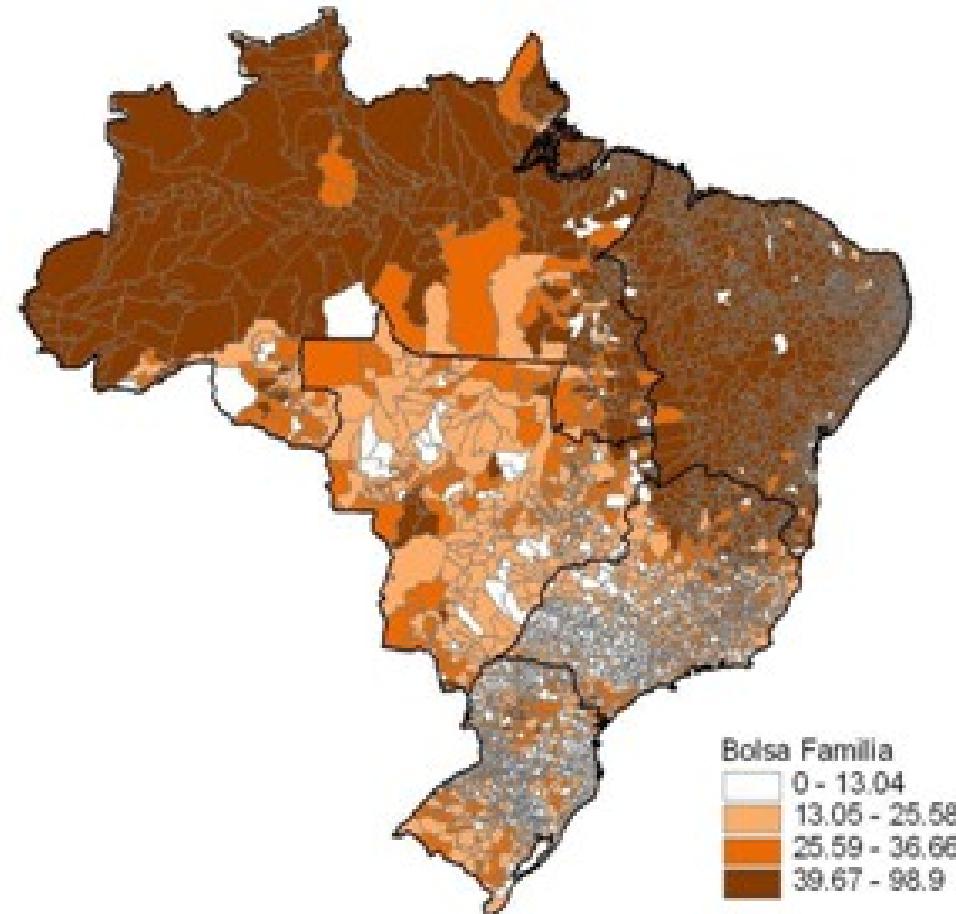
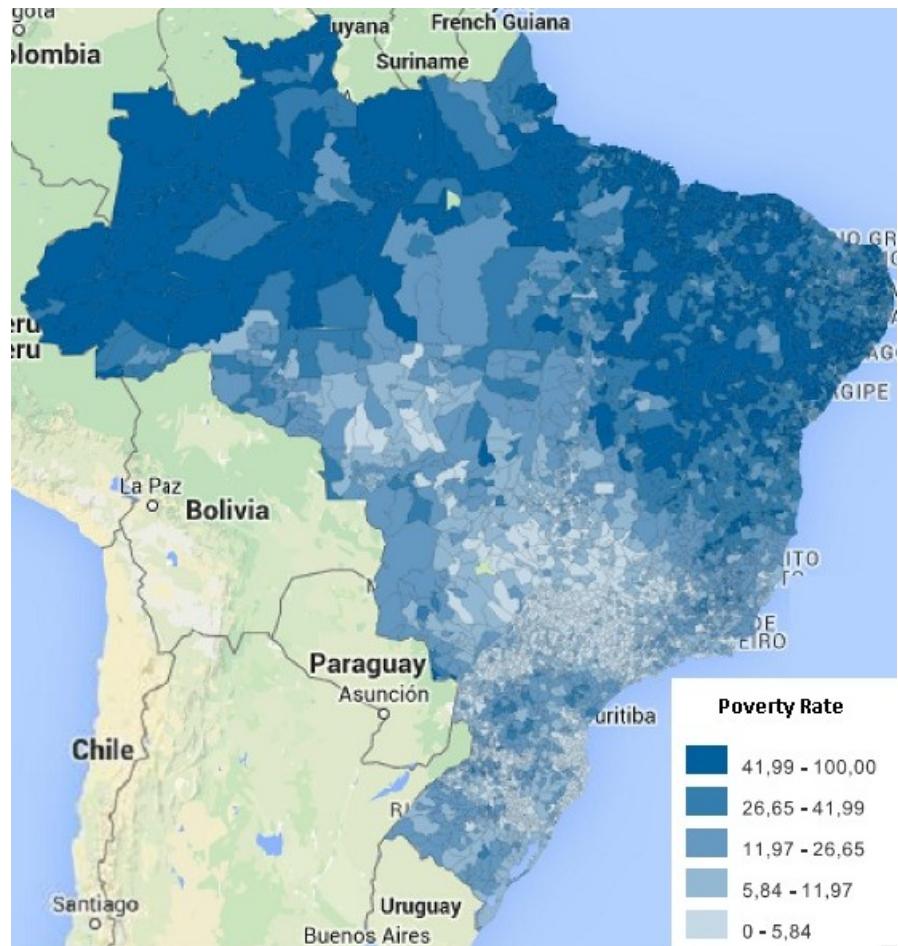


Social programmes

- **'Bolsa Família' Programme (BFP)**

- Launched in 2003 from four pre-existing social programmes.
- World's largest conditional cash transfer programme.
 - Extremely poor families (income of US\$ 35 per person per month).
 - Poor families (income between US\$ 35 and US\$ 70 per person / month).
 - US\$ 18 for each pregnant woman, child, and under-17 years adolescent.
 - Families must comply with education and health-related conditions.
- Main objectives:
 - promote access to public services (health, education and social assistance);
 - combat poverty and hunger, and promote safety food / nutrition;
 - stimulate sustained emancipation of families;
 - promote the synergy of the social actions taken by the government.

Poverty rate and Bolsa Família coverage



Programme	Reference	# of families	# of people	Full amount transferred BRL (US\$)
CadÚnico	March/2015	27,037,471	81,500,052	
BFP	April/2015	13,755,692		2,308,012,264.00 (770,982,183.32)

Effect of a conditional cash transfer programme on childhood mortality: a nationwide analysis of Brazilian municipalities

Davide Rasella, Rosana Aquino, Carlos A T Santos, Rômulo Paes-Sousa, Mauricio L Barreto

=> assessment of the effect of the BFP on deaths of children under-5 years, overall and resulting from poverty related diseases: malnutrition, diarrhea, and lower respiratory infections.

	2004	2005	2006	2007	2008	2009	Percentage change 2004-09
Mortality rate for children younger than 5 years (per 1000 livebirths)							
Overall	21.7 (14.7)	20.3 (14.5)	20.1 (14.6)	19.4 (14.8)	18.6 (15.9)	17.5 (14.7)	-19.4%
For diarrhoeal diseases	0.95 (2.93)	0.86 (2.54)	0.83 (2.67)	0.55 (2.02)	0.49 (1.96)	0.51 (2.46)	-46.3%
For malnutrition	0.55 (2.33)	0.48 (2.24)	0.36 (1.70)	0.30 (2.53)	0.20 (1.26)	0.23 (1.54)	-58.2%
For lower respiratory infections	1.15 (3.30)	0.96 (2.72)	1.07 (2.84)	0.95 (2.91)	0.98 (3.85)	0.84 (2.84)	-27.0%
For external causes	1.23 (3.29)	1.16 (3.14)	1.06 (3.17)	1.16 (3.80)	1.07 (3.70)	1.01 (3.71)	-17.9%
BFP coverage of the municipality population (%)	17.3% (12.1)	23.0% (14.0)	28.1% (17.2)	27.8% (17.8)	25.2% (16.7)	28.3% (17.5)	63.6%
FHP coverage of the municipality population (%)	62.7% (36.7)	67.8% (34.8)	71.0% (33.4)	73.9% (32.4)	74.4% (31.3)	75.0% (30.9)	19.6%
Income per person (monthly, in BR\$)	310 (126)	339 (135)	368 (145)	396 (154)	425 (164)	454 (147)	46.5%
Proportion of BFP eligible population in the municipality	27.9% (16.5)	27.8% (16.7)	27.8% (16.8)	27.7% (16.9)	26.5% (15.5)	26.3% (15.5)	-5.7%
Proportion of individuals living in households with inadequate sanitation	22.9% (16.4)	21.7% (15.8)	20.5% (15.2)	19.3% (14.7)	18.2% (14.3)	17.0% (13.9)	-25.8%
Proportion of individuals older than 15 years who are illiterate	16.9% (10.3)	16.4% (10.0)	15.9% (9.8)	15.4% (9.6)	14.9% (9.3)	14.4% (9.1)	-14.8%
Total fertility rate	2.31 (0.62)	2.27 (0.63)	2.20 (0.64)	2.14 (0.65)	2.07 (0.65)	2.01 (0.67)	-13.0%
Rate of admissions to hospital (per 100 inhabitants)	4.88 (4.47)	4.69 (4.34)	4.58 (4.39)	4.46 (4.11)	4.02 (4.11)	4.04 (4.23)	-17.2%

Data are mean (SD). Causes of death were defined according to the International Classification of Diseases, 10th revision:¹⁷ diarrhoeal diseases (A00, A01, A03, A04, A06–09), malnutrition (E40–46), lower respiratory infections (J10–18, J20–22), and external causes (V01–98). Rate of admission to hospital was calculated as the number of admissions to hospital for all ages and all causes of one municipality divided by the total population of the same municipality and multiplied by 100. BFP=Bolsa Família Programme. FHP=Family Health Programme.

Table 1: Mortality rates and variables for selected municipalities (N=2853)

Effect of the Brazilian Conditional Cash Transfer and Primary Health Care Programs on the New Case Detection Rate of Leprosy

Joilda Silva Nery^{1*}, Susan Martins Pereira¹, Davide Rasella¹, Maria Lúcia Fernandes Penna², Rosana Aquino¹, Laura Cunha Rodrigues³, Mauricio Lima Barreto¹, Gerson Oliveira Penna⁴

PLOS Neglected Tropical Diseases | www.plosntds.org November 2014 | Volume 8 | Issue 11 | e3357

Table 1. Number of new cases and new case detection rate of leprosy in the Brazil and selected municipalities (n = 1,358), Brazil 2004–2011.

Year	Number of new cases - Selected municipalities (a)	Total number of new cases -Brazil (b)	% of cases the total of Brazil (a/b)	Leprosy new case annual detection rate* - Selected municipalities	Leprosy new case annual detection rate* - Brazil
2004	30,024	50,565	59.3	74.8	28.2
2005	29,740	49,448	60.1	73.0	26.8
2006	26,908	43,642	61.6	65.1	23.3
2007	25,165	40,126	61.7	61.5	21.1
2008	24,816	39,047	63.5	58.8	20.5
2009	22,943	37,610	61.0	53.7	19.6
2010	21,469	34,894	61.5	49.8	18.2
2011	19,901	33,955	58.6	45.6	17.6

*Per 100,000 inhabitants.

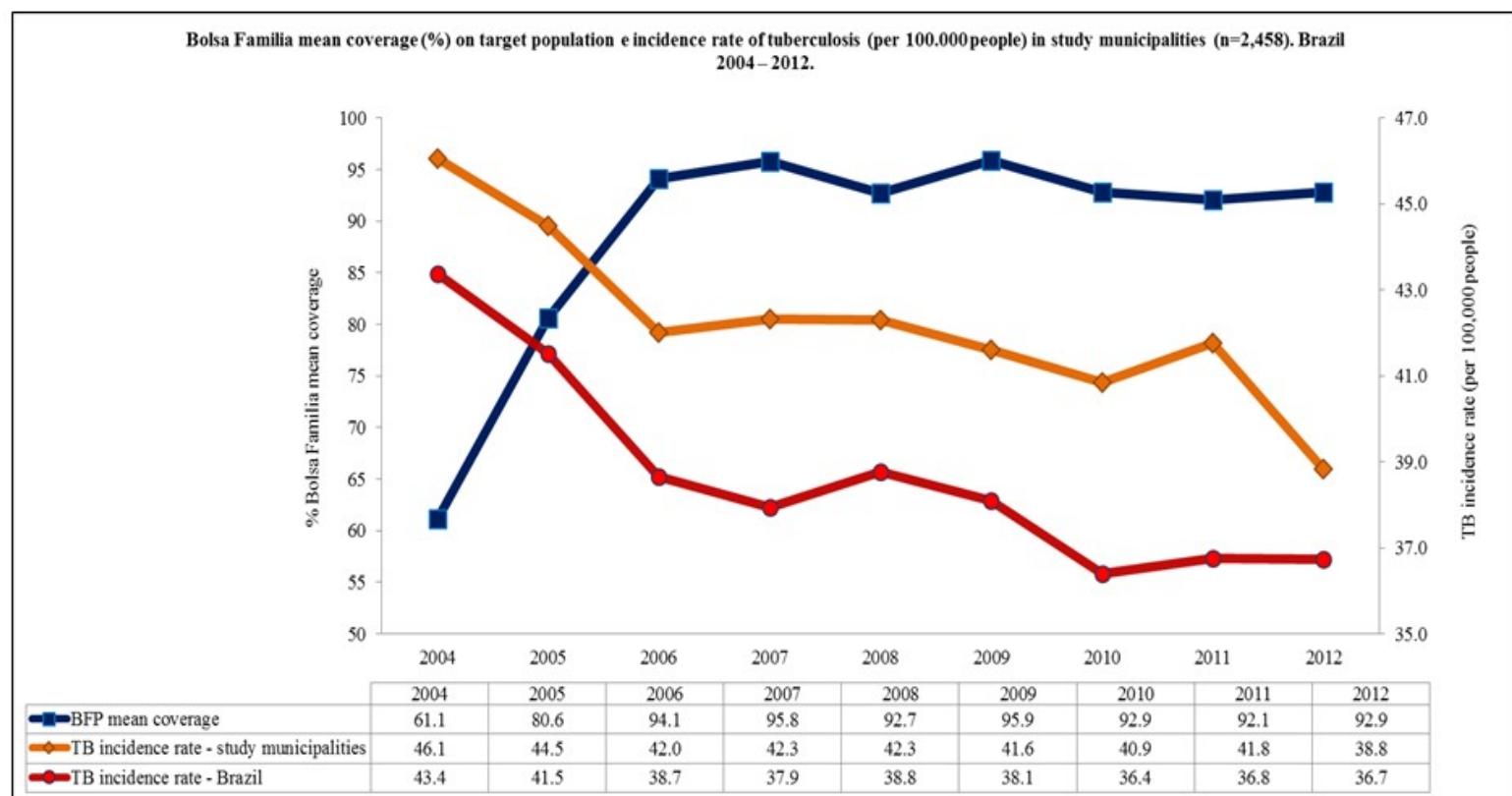
doi:10.1371/journal.pntd.0003357.t001

Concluding this part...

- Conditional cash transfer (as BFP) + primary care (as FHP) can effectively contribute greatly to the reduction of poverty-related diseases.

Figure 1: BolsaFamília mean coverage (%) on target population e incidence rate of tuberculosis (per 100.000 people) in study municipalities (n=2,458). Brazil 2004 – 2012.

Another example:
Tuberculosis



- Prevalence of mixed-ecological studies (ecological multi-group + time-trend) having municipalities as unit of analysis.

Outline

- Brazilian health system and social programmes
- **Project overview**
- Spark-based pipeline for record linkage
- Cohort design and first steps

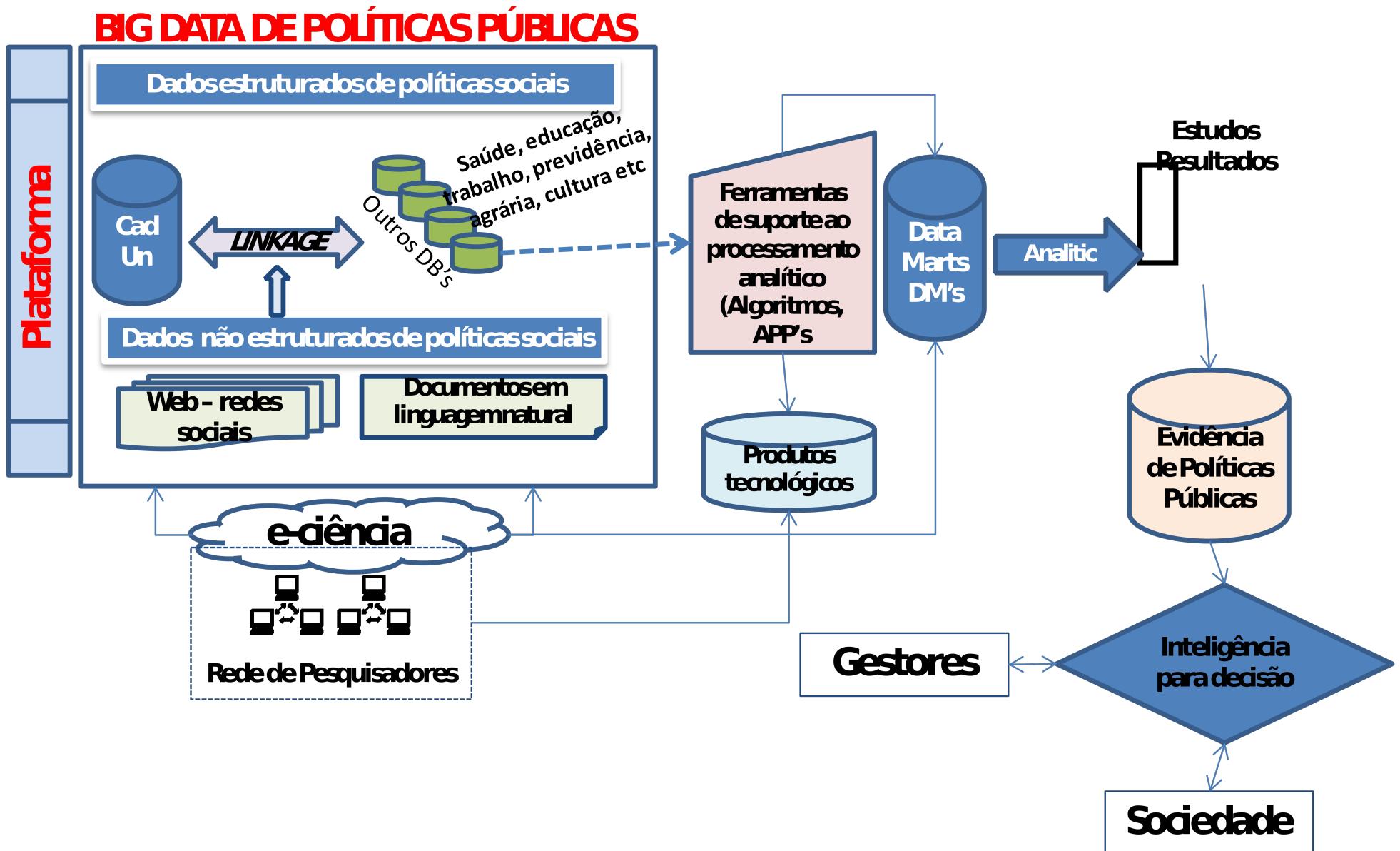
Main project

- Maurício Barreto
 - A platform for continuous studies and assessments of the effects of the Bolsa Família Programme and other social protection programmes on health, education, labor and race/gender relations based on a populational cohort from Cadastro Único.
- Related projects
 - Davide Rasella
 - Effects on tuberculosis and leprosy; predictive analysis.
 - Maria Yuri
 - Effects on HIV/AIDS; currently working on livebirth conditions X prenatal visits.
 - Rosemeire Fiaccione and Leila Amorim
 - Special analytical methods (Regression Discontinuity Design, Propensity Score Matching etc) applied to leprosy and tuberculosis in the state of Bahia.
 - Marcos Barreto
 - Computing platform to support such projects.

Aims

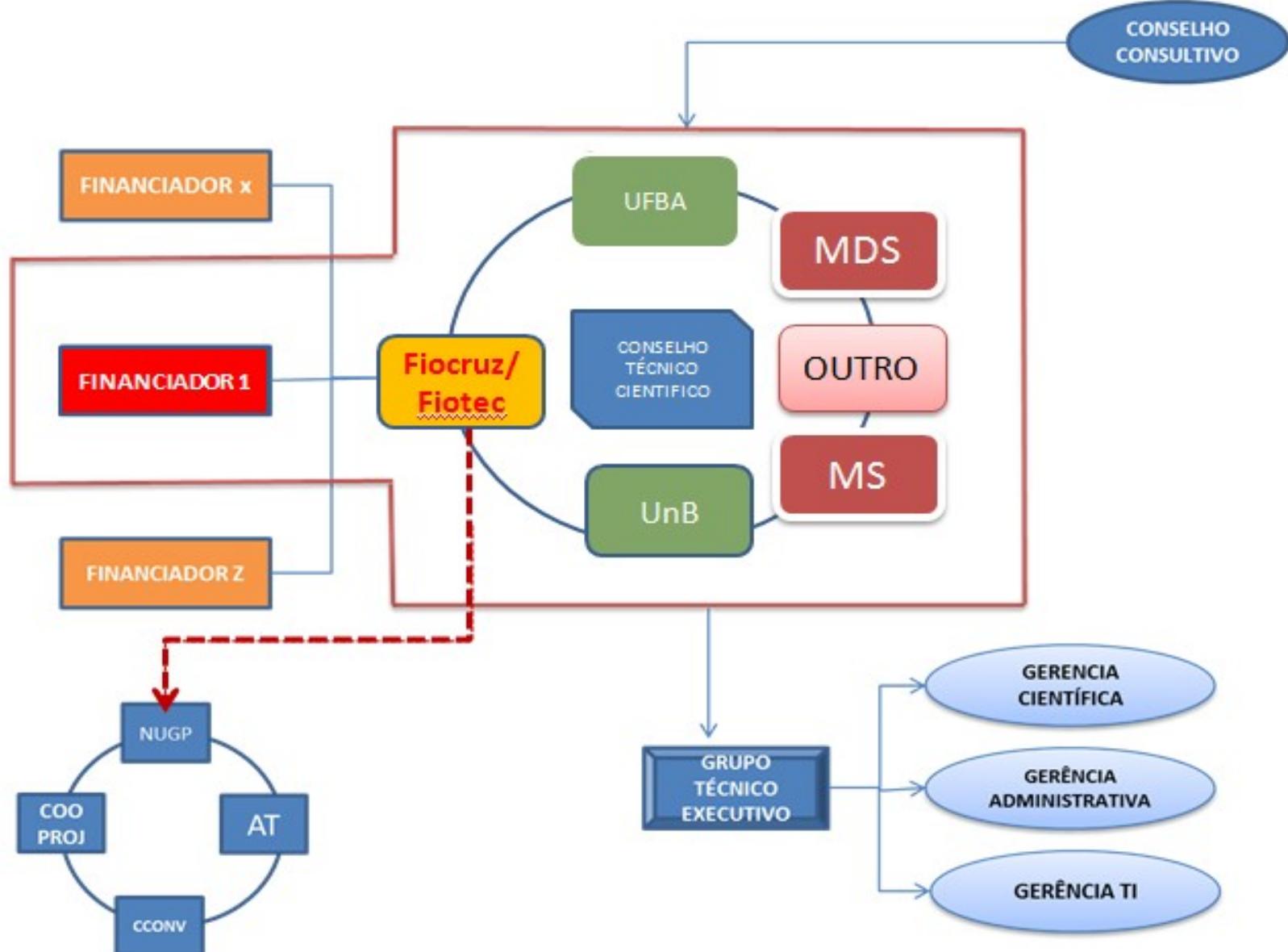
- Develop a population-based cohort from Cadastro Único and progressively accumulate information from other routine databases to be added through processes of linkage.
- This tool would aim to answer questions related to:
 - assessment of the effects (on health, education, labor, gender relations etc.) of Bolsa Familia and other social programmes;
 - studies on social determinants of health, education, work etc;
 - the establishment of procedures and mechanisms to enable access to data resources and analysis for managers and researchers with policy, managerial or scientific questions.

Project workflow



From Maurício's project proposal

Project management



From Maurício's project proposal

Databases



Ministério do
Desenvolvimento Social
e Combate à Fome

- ✗ Very disparate databases without reliable common attributes.
- ✗ Need for probabilistic record linkage methods.

Databases	Years
SIH (hospitalization)	1998 to 2011
SINAN (notifiable diseases)	2000 to 2012
SIM (mortality)	2000 to 2012
CadÚnico (socioeconomic data)	2007 to 2013
PBF (payments from Bolsa Família)	2007 to 2013

- ✗ Deterministic linkage through the **NIS** (social ID number) attribute.
- ✗ Persons registers must be renewed each two years.

=> PBF runs each month; the other databases run once a year.

Outline

- Brazilian health system and social programmes
- Project overview
- **Spark-based pipeline for record linkage**
- Cohort design and first steps

Spark

- Designed by Databricks, based on the MapReduce programming model and the Hadoop framework.
- Main features:
 - Compatible with general-purpose languages (Java, Python, and Scala).
 - RDD (resilient distributed dataset).
 - Fault tolerant collection of partitionable and distributable objects.
 - Manipulation based on **transformations** and **actions**. (*next slide*)
 - Parameterized allocation (MEMORY_ONLY, MEMORY_AND_DISK, MEMORY_ONLY_SER, MEMORY_AND_DISK_SER, DISK_ONLY).
 - In-memory data processing.
 - Usage of **broadcast variables** and **accumulators**.
 - Storage of temporary/partial results within iterative computations.

Spark

- RDD Application Programming Interface (API)

Transformations

Generate a new dataset from an existing one.

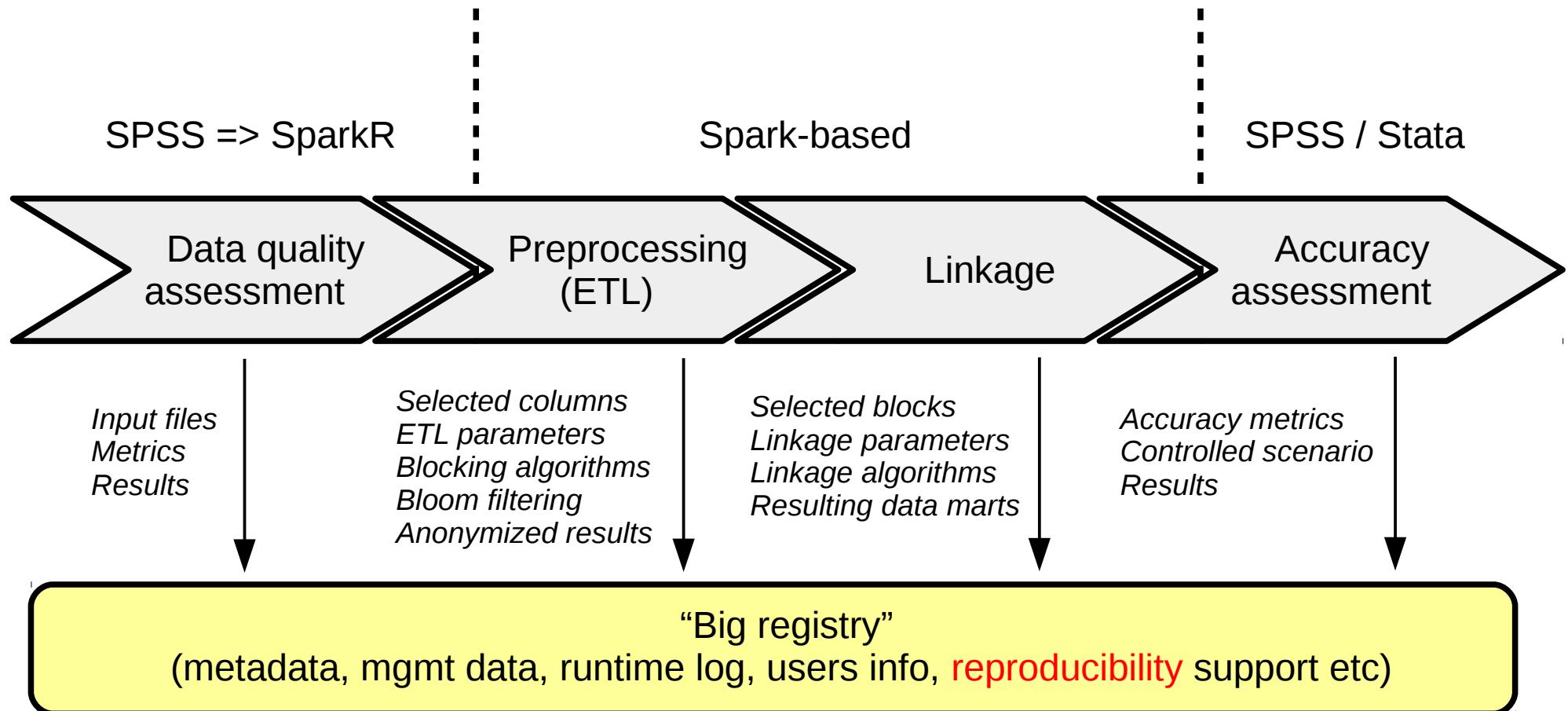
- x `map(func)`
- x `flatMap(func)`
- x `filter(func)`
- x `distinct([num_tasks])`
- x `reduceByKey(func, [num_tasks])`

Actions

Return values to the program after processing a RDD

- x `collect()`
- x `count()`
- x `foreach(func)`
- x `reduce(func)`

Pipeline overview



Case study #1 – CadÚnico x SIH (year 2011)

- Our first implementation!
- Data quality assessment

A Spark-based workflow for probabilistic record linkage of healthcare data (BeyondMR 2015)

Attribute	Description	Missing (%)
NIS	Social number ID	0,7
NOME	Person's name	0
DT_NASC	Date of birth	0
MUNIC_RES	City of residence	0
SEXO	Gender	0
RG	General ID	48,7
CPF	Individual taxpayer ID	52,1

CadÚnico

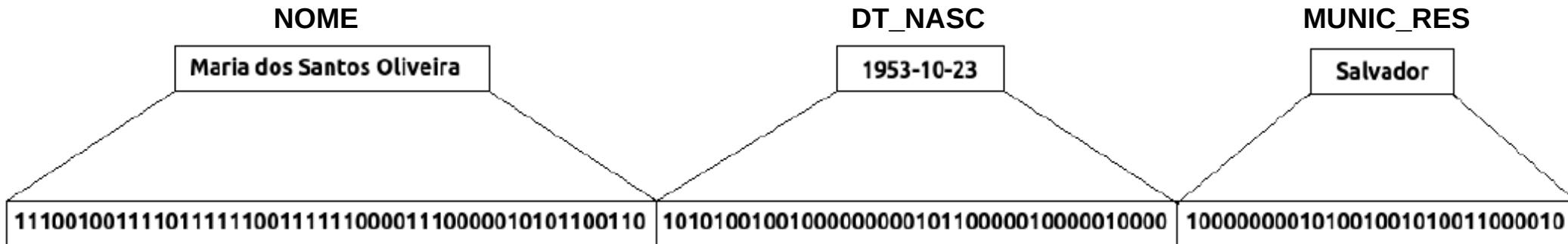
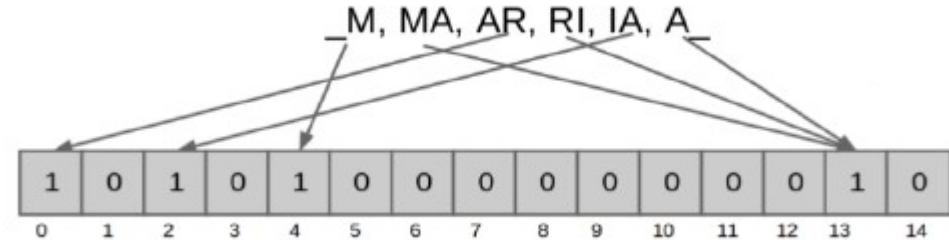
Chosen attributes:
- NOME
- DT_NASC
- MUNIC_RES

Attribute	Description	Missing (%)
MUNIC_RES	City of residence	0
NASC	Date of birth	0
SEXO	Gender	0
NOME	Patient's name	0
LOGR	Street name	0,9
NUM_LOGR	House number	16,4
COMPL_LOGR	Address' complement	80,7

SIH

Case study #1

- Pre-processing (ETL)
 - Bloom filtering



Decision on vector size and weights...

Total size and weight distribution	Scenario 1 No matched records expected		Scenario 2 Five perfectly matched records expected		Scenario 3 Expected five matched records with one incorrect character	
	Expected pairings	Pairings found	Expected pairings	Pairings found	Expected pairings	Pairings found
20x20x20	0	310	5	347	5	348
30x30x30	0	29	5	41	5	42
40x40x40	0	11	5	17	5	16
50x50x50	0	0	5	5	5	5
50x50x40	0	0	5	5	5	5
50x40x40	0	0	5	5	5	5
50x40x30	0	0	5	5	5	5
50x30x30	0	2	5	6	5	6
50x40x20	0	0	5	5	5	5

110 bits

Case study #1 – CadÚnico x SIH (year 2011)

- Pre-processing (ETL)
 - **RDD manipulation**

Transformation	Meaning
map(func)	Returns a new RDD by passing each element of the source through <i>func</i>
mapPartitions(func)	Similar to map, but runs separately on each partition (block) of the RDD.
Action	Meaning
collect()	Returns all the elements of the dataset as an array at the driver program.
count()	Returns the number of elements in the dataset.

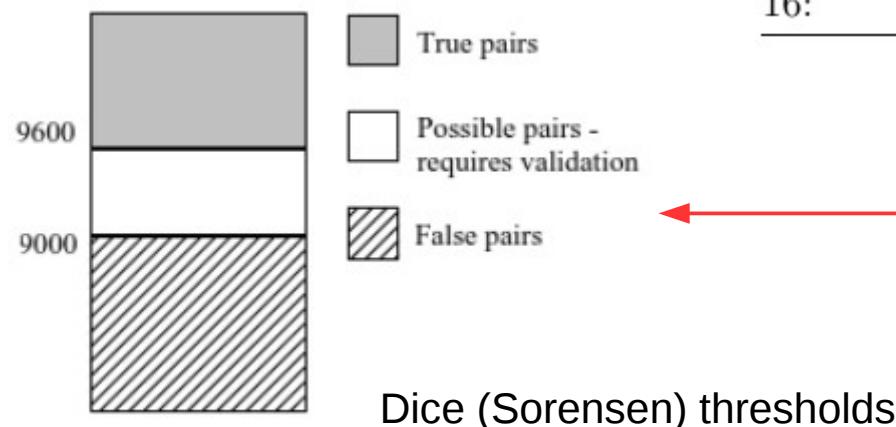
Algorithm 1 PreProcessing

```
1: Input ← OriginalDatabase.csv
2: Output ← TreatedDatabaseAnom.bloom
3: InputSparkC ← sc.textFile(Input)
4: NameSize ← 50
5: BirthSize ← 40
6: CitySize ← 20
7: ResultBeta ← InputSparkC.cache().map(normalize)
8: Result ← ResultBeta.cache().map(blocking).collect()
9: for line in Result:
10:   write line in Output
11: procedure NORMALIZE(rawLine)
12:   splitedLine ← rawLine.split(;)
13:   for fields in splitedLine:
14:     field ← field.normalized(UTF8) return splitedLine.join(;)
15: procedure BLOCKING(treatedLine)
16:   splLine ← treatedLine.split(;)
17:   splLine[0] ← applyBloom(splLine[0], NameSize)
18:   splLine[1] ← applyBloom(splLine[1], BirthSize)
19:   splLine[2] ← applyBloom(splLine[2], CitySize)
    return splitedLine.join()
20: procedure APPLYBLOOM(field, vectorSize)
21:   instanceInitialVectorWithSize ← vectorSize
22:   for n-grams in field:
23:     bitsVector ← Calculate positions of 1s in Vector
    return bitsVector
```

Case study #1 – CadÚnico x SIH (year 2011)

- Linkage
 - RDD manipulation

Transformation	Meaning
map(func)	Returns a new RDD by passing each element of the source through <i>func</i>
mapPartitions(func)	Similar to map, but runs separately on each partition (block) of the RDD.
Action	Meaning
collect()	Returns all the elements of the dataset as an array at the driver program.
count()	Returns the number of elements in the dataset.



Algorithm 2 Record linkage

```
1: InputMinor ← TreatedDatabaseAnom1.bloom
2: InputLarger ← TreatedDatabaseAnom2.bloom
3: InputSC1 ← sc.textFile(InputMinor)
4: InputSC2 ← sc.textFile(InputLarger)
5: var ← InputSC1.cache().collect()
6: varbc ← sc.broadcast(var)
7: InterResult ← InputSC2.cache().map(compare)
8: Result ← InputSC2.cache().collect()
9: for line in recordLinkageResult:
10: write line in Output
11: procedure COMPARE(line)
12: for linebc in varbc.value:
13:     get Dice index of (linebc) and (line) comparison
14:     decide about the similarity
15:     if Dice = 9000 then return line
16: else return None
```

Case study #1 – CadÚnico x SIH (year 2011)

- Linkage => blocking by state
 - 3 samples (smallest Brazilian states)

Sample Name	Size (in lines) CadÚnico x SIH	Comparisons (millions)	Exec. Time (seconds)
A	367,892 x 147	54,0	96,26
B	1,6 mi x 171	289,5	479
C	1,02 mi x 389	397,63	656,79

Tools	i5	i7	Cluster
Spark	507.5 s	235.7 s	96.26 s
OpenMP	104.9 s	65.5 s	13.36 s

- Whole database

	CadÚnico	SIH
Size (lines)	approx. 87 mi	approx. 61 k
Standardization		
Anonymization	2310.4 s	36.5 s
Blocking		
Record Linkage	9,03 hours	
Paired Recovery	1,31 hours	

Case study #2 – CadÚnico x SIH x SIM (year 2011)

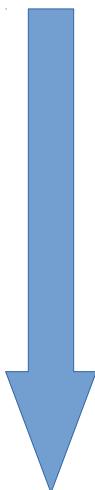
Databases	Records
CadÚnico	~ 106,000,000
SIH	~ 62,000
SIM	~ 17,000

- Data quality assessment

CadÚnico	
Atributo	Missing %
NOME	0
DT_NASC	0
MUNIC_RES	55,4
SEXO	0
RG	48,7

SIH	
Atributo	Missing %
NOME	0
NASC	0
MUNIC_RES	0
SEXO	0
LOGR	0,9

SIM	
Atributo	Missing %
NOME	9
NASC	1,2
RES	0



- ETL (deduplication)

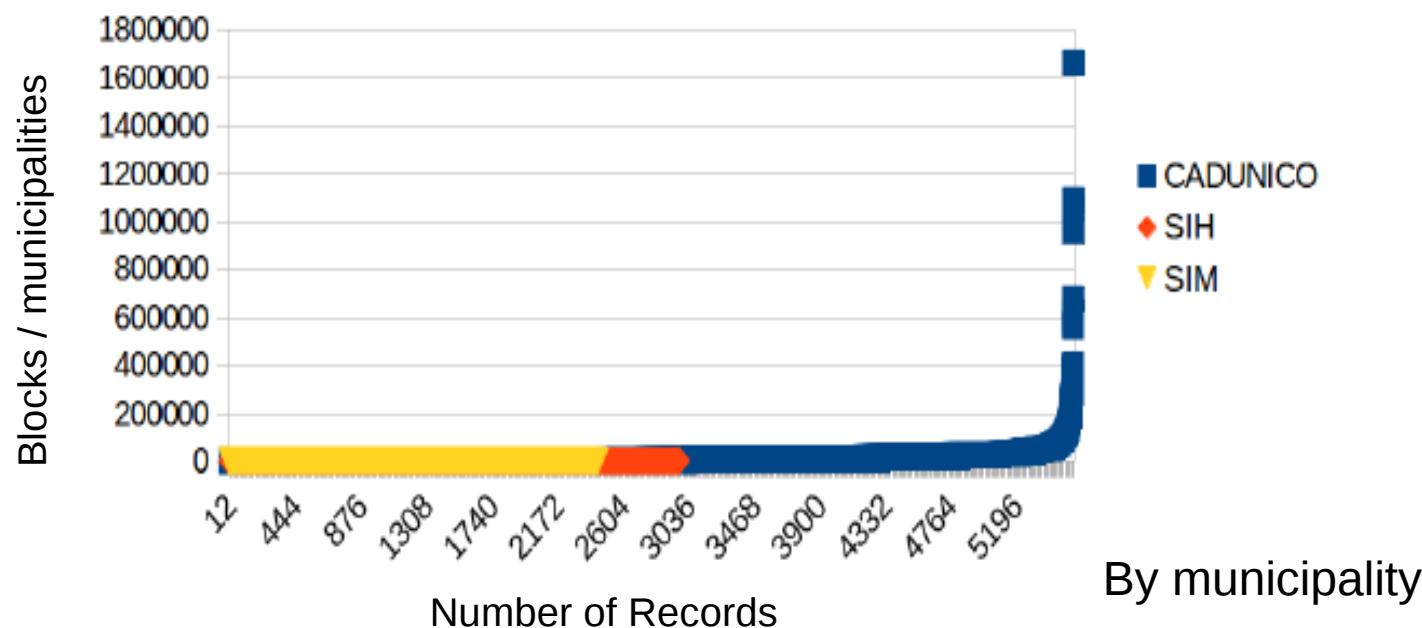
Databases	Records
CadÚnico	~ 76,000,000
SIH	~ 56,000
SIM	~ 17,000

Case study #2 – CadÚnico x SIH x SIM (year 2011)

- ETL (blocking)

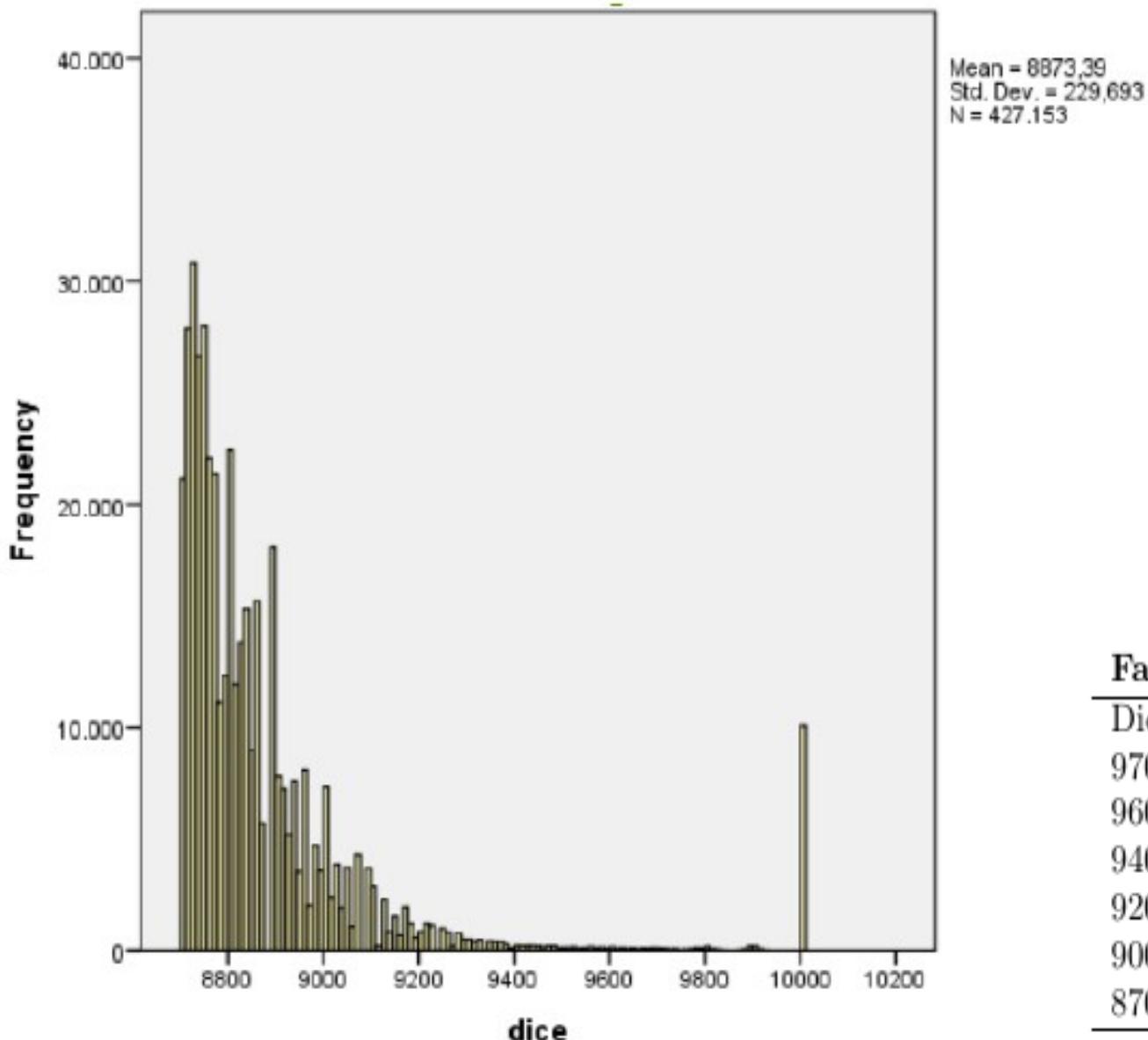
By state

Estado	CadÚnico	SIH	Estado	CadÚnico	SIH
	Tam. (%)	Tam. (%)		Tam. (%)	Tam. (%)
Rondônia	0,8	1,1	Sergipe	1,7	0,5
Acre	0,5	0,2	Bahia	11,0	4,6
Amazonas	0,2	2	Minas Gerais	9,7	6,4
Roraima	0,4	0,3	Esp. Santo	1,8	1,4
Amapá	0,4	0,2	R. de Janeiro	4,9	11,2
Tocantins	1,0	0,6	São Paulo	13,0	25,4
Maranhão	6,1	1,6	Paraná	5,2	3,5
Piauí	3,2	0,9	St. Catarina	1,9	4,8
Ceará	8,0	3,5	R. G. do Sul	4,2	11,8
Rio G. do Norte	2,6	1,7	Mt. Gr. Sul	2,2	1,5
Paraíba	3,5	3,6	Mato Grosso	2,3	0,8
Pernambuco	7,3	7,8	Goiás	3,2	2,9
Alagoas	3,1	1,2	Dist. Federal	1,8	0,5



Case study #2 – CadÚnico x SIH x SIM (year 2011)

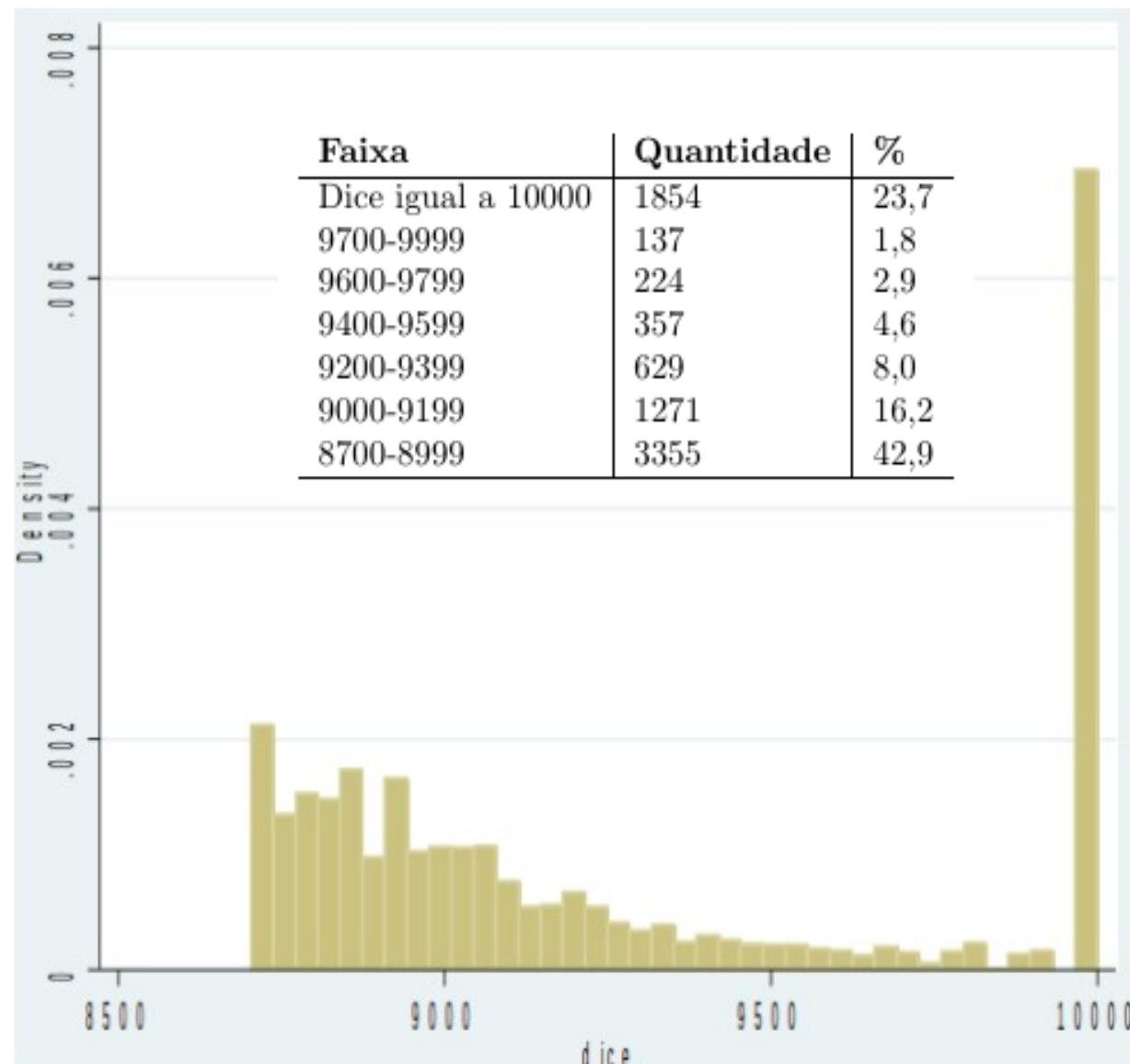
- Linkage: CadÚnico x SIH (entire database)



Faixa	Quantidade	%
Dice igual a 10000	10089	2,4
9700-9999	651	0,2
9600-9799	1277	0,3
9400-9599	2502	0,6
9200-9399	9273	2,2
9000-9199	40252	9,4
8700-8999	363109	85,0

Case study #2 – CadÚnico x SIH x SIM (year 2011)

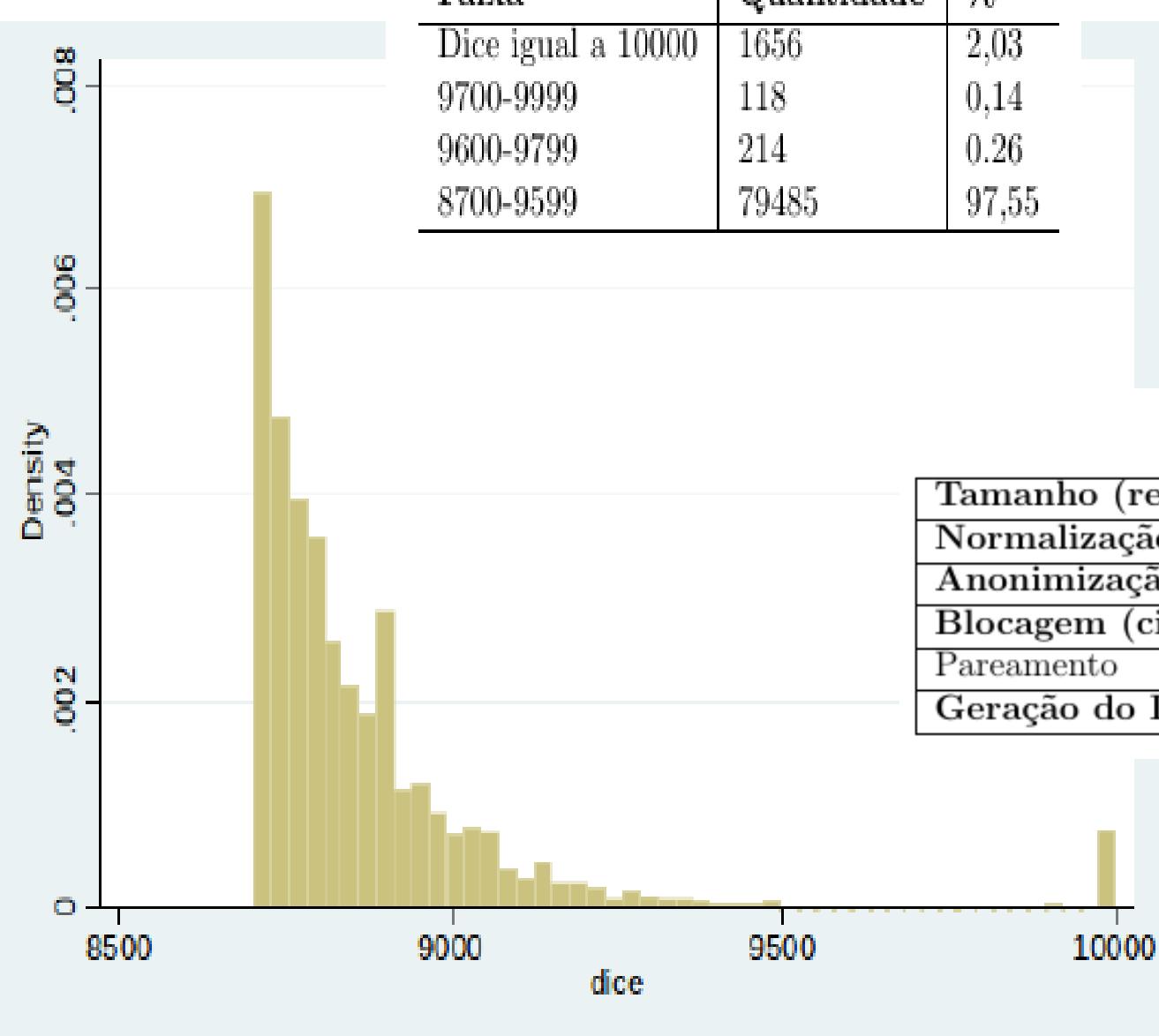
- Linkage: CadÚnico x SIH (tuberculosis)



CadÚnico x SIH_TB		
Tamanho (registros)	~76 milhões	~15 mil
Normalização (UTF8)		
Anonimização (Bloom)	2310,4 s	7,9 s
Blocagem (cidades)		
Pareamento	2,9 h	
Geração do DataMart	1268,98 s	

Case study #2 – CadÚnico x SIH x SIM (year 2011)

- Linkage: CadÚnico x SIM (tuberculosis)



CadÚnico x SIM_TB	
~76 milhões	~17 mil
Normalização (UTF8)	
Anonimização (Bloom)	
Blocagem (cidades)	2310,4 s
Pareamento	8,3 s
Geração do DataMart	3,6 h
	456 s

Case study #3 – Accuracy assessment

- Diarrhea (rotavirus) database
 - 486 records of infected children X 9,678 records of children hospitalized for diarrhea and other diseases.
 - Four evaluation scenarios:
 - Modifications in NOME and DATA_NASC attributes:
 - character and digit replacement, surname exchange.
 - Scenario 1: 10.3% records changed.
 - Scenario 2: 11.3% records changed.
 - Scenario 3: 10.3% records changed.
 - Scenario 4: 5.15% records changed.

- x Without blocking
- x 100-bit Bloom
- x Dice considers the entire Bloom filter

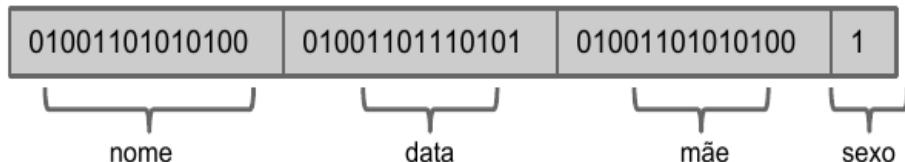
	Scenario 1	Scenario 2	Scenario 3	Scenario 4
True matches	482	481	479	482
True non-matches	9	7	7	7
Missing pairs	4	5	7	4

Dice	Scenario 1			Scenario 2		
	Sensitivity (%)	Specificity (%)	VPP (%)	Sensitivity (%)	Specificity (%)	VPP (%)
≥10000	8,8	100,0	100,0	42,8	100,0	100,0
≥9800	12,8	100,0	100,0	52,7	100,0	100,0
≥9600	59,5	100,0	100,0	80,0	100,0	100,0
≥9400	86,6	100,0	100,0	94,4	100,0	100,0
≥9200	95,3	100,0	100,0	97,1	100,0	100,0
≥9000	98,1	100,0	100,0	98,4	100,0	100,0
≥8800	98,8	100,0	100,0	98,8	99,0	0,996
≥8600	99,0	100,0	100,0	99,0	96,5	0,986
≥8400	99,2	99,5	99,8	99,0	96,5	0,986
≥8200	99,2	99,5	99,8	99,0	96,5	0,986
≥8000	99,2	99,5	99,8	99,0	96,5	0,986
≥7000	99,2	95,5	98,2	99,0	96,5	0,986
Dice	Scenario 3			Scenario 4		
	Sensitivity (%)	Specificity (%)	VPP (%)	Sensitivity (%)	Specificity (%)	VPP (%)
≥10000	42,2	100,0	100,0	42,8	100,0	100,0
≥9800	51,9	100,0	100,0	53,1	100,0	100,0
≥9600	78,2	100,0	100,0	81,5	100,0	100,0
≥9400	93,4	100,0	100,0	96,1	100,0	100,0
≥9200	97,5	100,0	100,0	98,8	100,0	100,0
≥9000	98,4	100,0	100,0	99,0	100,0	100,0
≥8800	98,6	99,0	99,6	99,2	99,0	99,6
≥8600	98,6	96,5	98,6	99,2	96,5	98,6
≥8400	98,6	96,5	98,6	99,2	96,5	98,6
≥8200	98,6	96,5	98,6	99,2	96,5	98,6
≥8000	98,6	96,5	98,6	99,2	96,5	98,6
≥7000	99,2	95,5	98,2	99,0	96,5	0,986

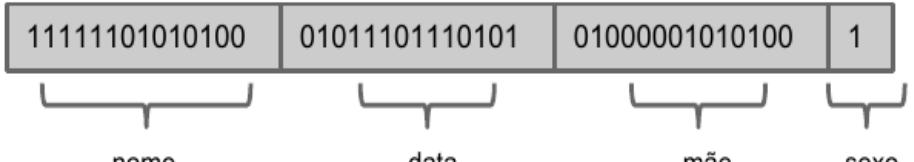
Case study #3 – Accuracy assessment

- Without blocking
- 100-bit Bloom
- Isolated comparison of each attribute

registro1



registro1



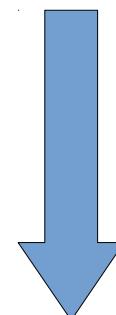
	Unmodified scenario	Scenario 1	Scenario 2	Scenario 3	Scenario 4
True matches	486	482	482	480	486
True non-matches	0	0	0	0	0
Missing pairs	0	4	4	6	0

Case study #3 – Accuracy assessment

- ✗ **Multiple blocking (*blocagem por predicados*)**
- ✗ (NOME and MUNIC_RES) OR (SURNAME and DATA_NASC)

	Unmodified scenario	Scenario 1	Scenario 2	Scenario 3	Scenario 4
True matches	486	469	413	468	466
True non-matches	0	0	0	0	0
Missing pairs	0	17	73	18	20

Codificação SOUNDEX	
Código fonético	Letras
1	B,F,P,V
2	C, G, J, K, Q,S, X, Z—
3	D, T
4	L
5	M, N
6	R



Enter phonetic codes!

Codificação BuscaBR			
Fonema	Letras	Fonema	Letras
I	I	M	N, RM, GM, MD, SM e terminação em AO
B	BR	N	NH
F	PH	P	P
G	GR, MG, NG, RG	S	Ç, X, TS, C, Z, RS
J	GE, GI, RJ, MJ, NJ	T	LT, TR, CT, RT, ST
K	Q, CA, CO, CU, C	V	W
L	LH	L	R

Codificação Metaphone_pt-BR			
Fonema	Letras	Fonema	Letras
R	CR, *R*,	Z	EX*
K	Q, CA, CO, CU, C, K, CHR	T	TH, T, T*
S	CE, CI, Ç, SS, SCE, SCI	F	PH
G	GA, GO, GU, GH	L	L*
M	M, N*,	2	R, R\$, RR
J	GE, GI, GHE, GHI	3	NH
X	SCH, SH, CH, EXE, EXI,	1	R

Case study #4 – Phonetic codes

	SOUNDEX	BUSCABR	METAPHONE-PT_BR
Kubitscheck	K132	KBSK	KBTXK
Kubixequ	K122	KBSK	KBXK
Walter	W436	VT	VLTR
Valter	V436	VT	VLTR
Teresina	T625	TRSM	TRZN
Terezina	T625	TRSM	TRZN

	Edit distance	Jaro-Winkler	Jaccard	Fuzzy String Matching (FSM)	Sorensen (Dice)
Kubitscheck x Kubieque	6	0.6700	0.5833	0.4	0.58
Walter x Valter	1	0.8888	0.2857	0.83	0.83
Teresina x Terezina	1	0.95	0.25	0.88	0.85

Case study #5 – Comparison with other tools

- Two small databases
 - 126 records x 73 records

	Frill	Merge toolbox (German RLC)	Spark
Total matches	69	73	73
False positive	0	1	0
False negative	4	1	0

Outline

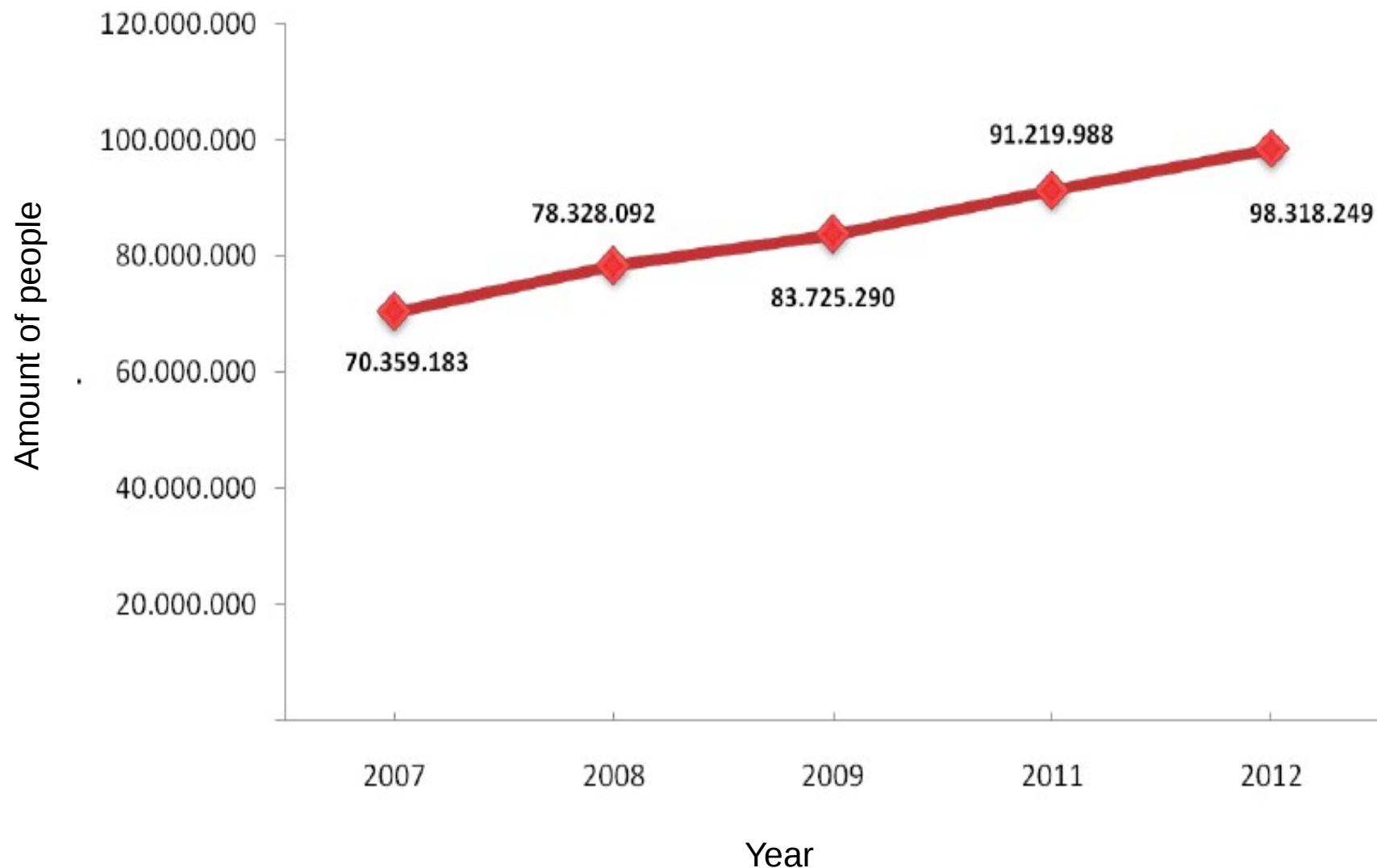
- Brazilian health system and social programmes
- Project overview
- Spark-based pipeline for record linkage
- Cohort design and first steps

Aims and relevance

- Develop a population-based cohort from the Cadastro Único database to continuously evaluate the effectiveness of the social programmes on health, education etc.
 - Coverage of the BFP's beneficiaries directly associated with the studied diseases, exposure time, and how much time and amount are necessary to impact health.
- Epidemiological standpoint: a way to evaluate social public programmes on different areas.
- Governmental standpoint: a way to monitor and improve such programmes.
- Computing standpoint: several issues regarding performance, accuracy, and big data processing.

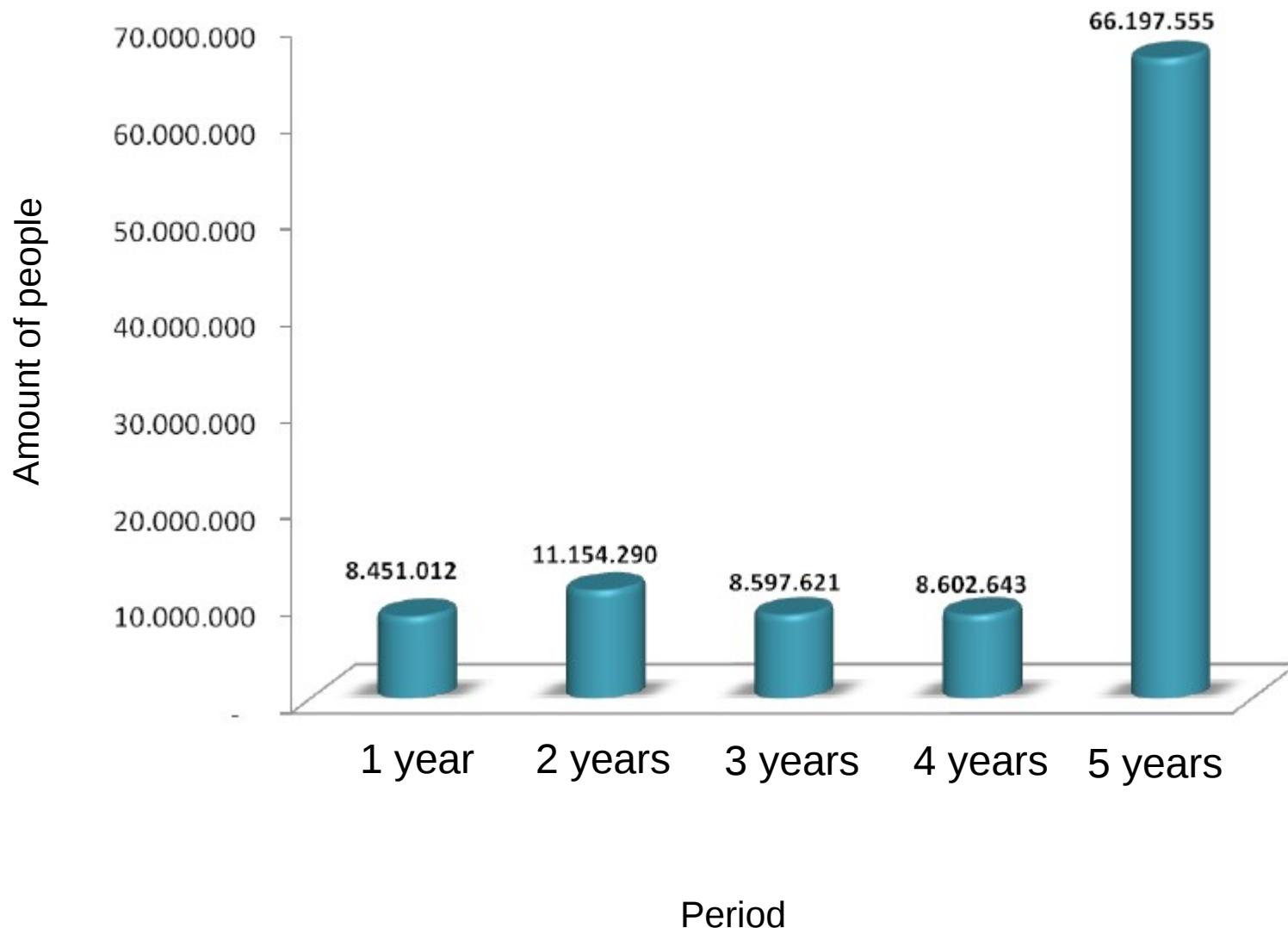
Cohort characterization

- People registered in CadÚnico



Cohort characterization

- People registered in CadÚnico by year



Cohort characterization

- People with at least one participation (basis 2007)



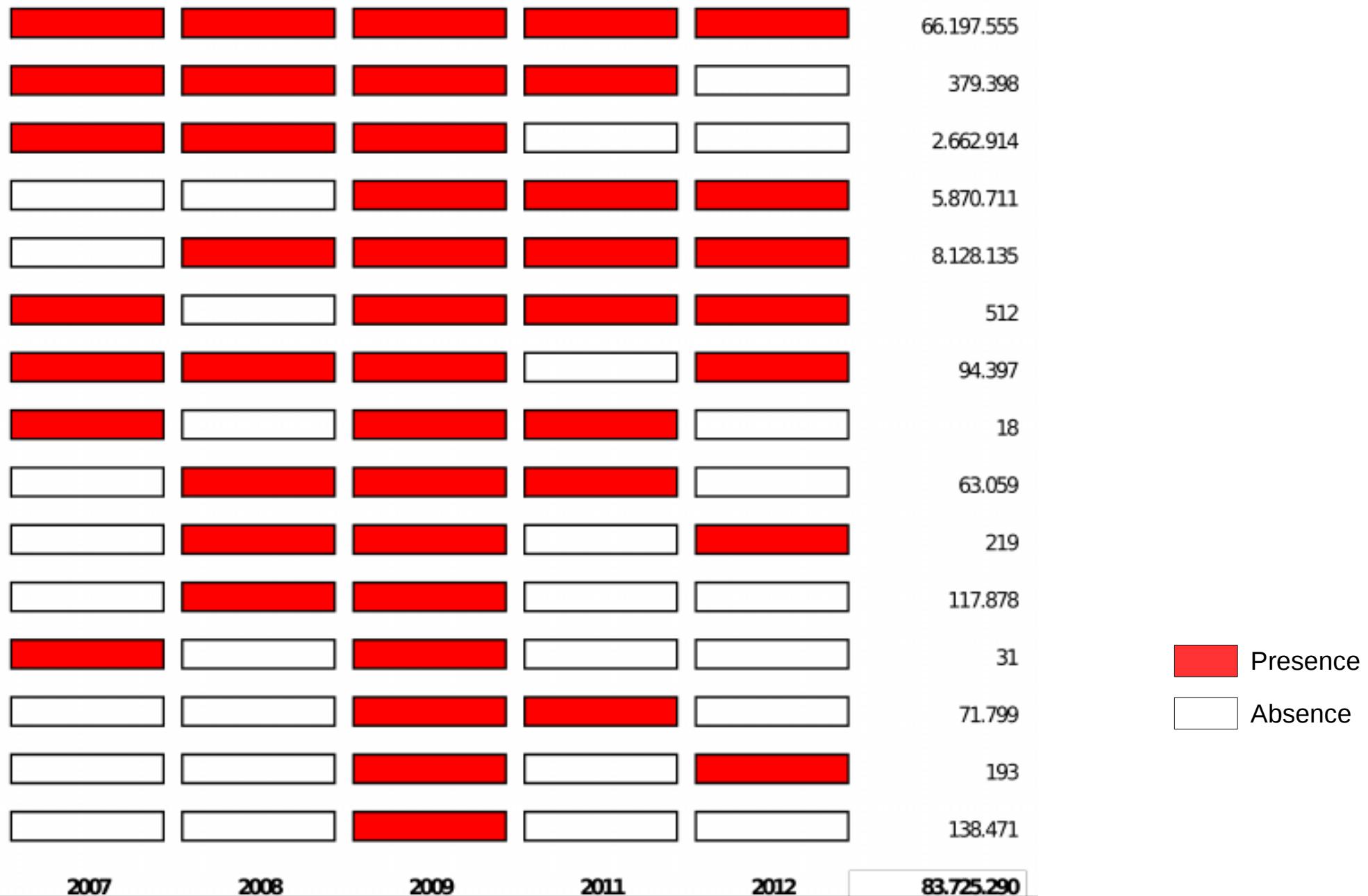
Cohort characterization

- People with at least one participation (basis 2008)



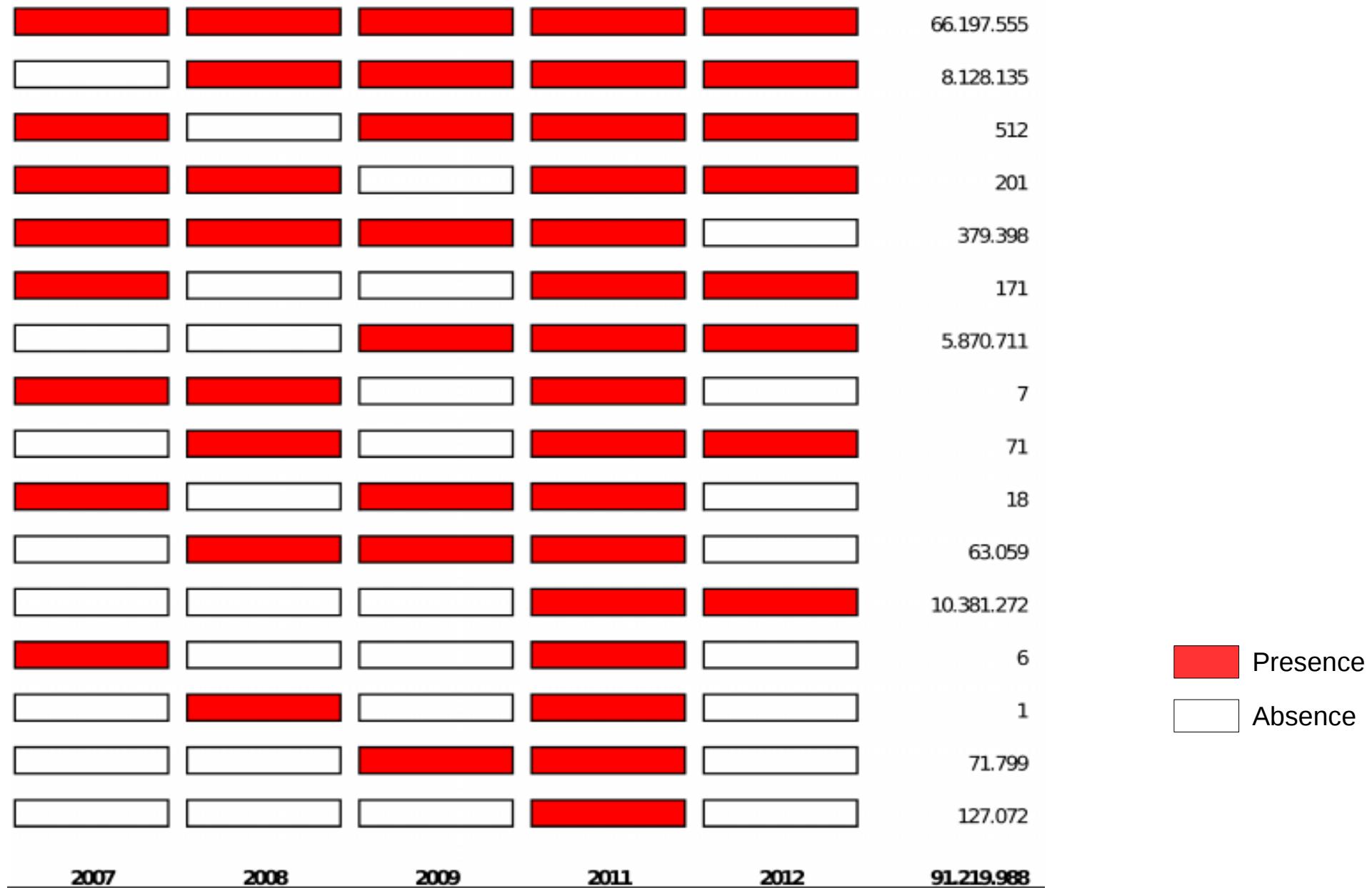
Cohort characterization

- People with at least one participation (basis 2009)



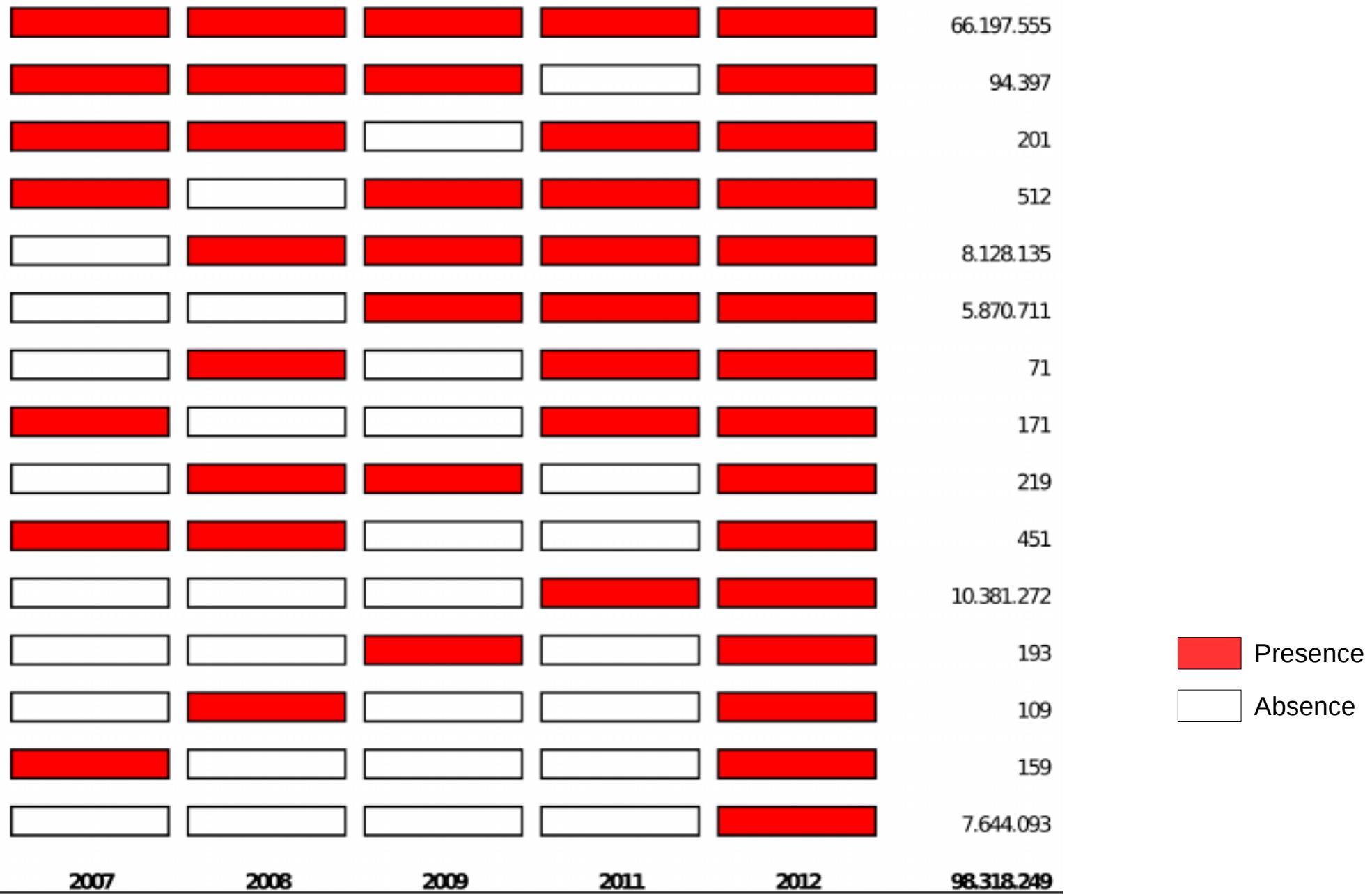
Cohort characterization

- People with at least one participation (basis 2011)



Cohort characterization

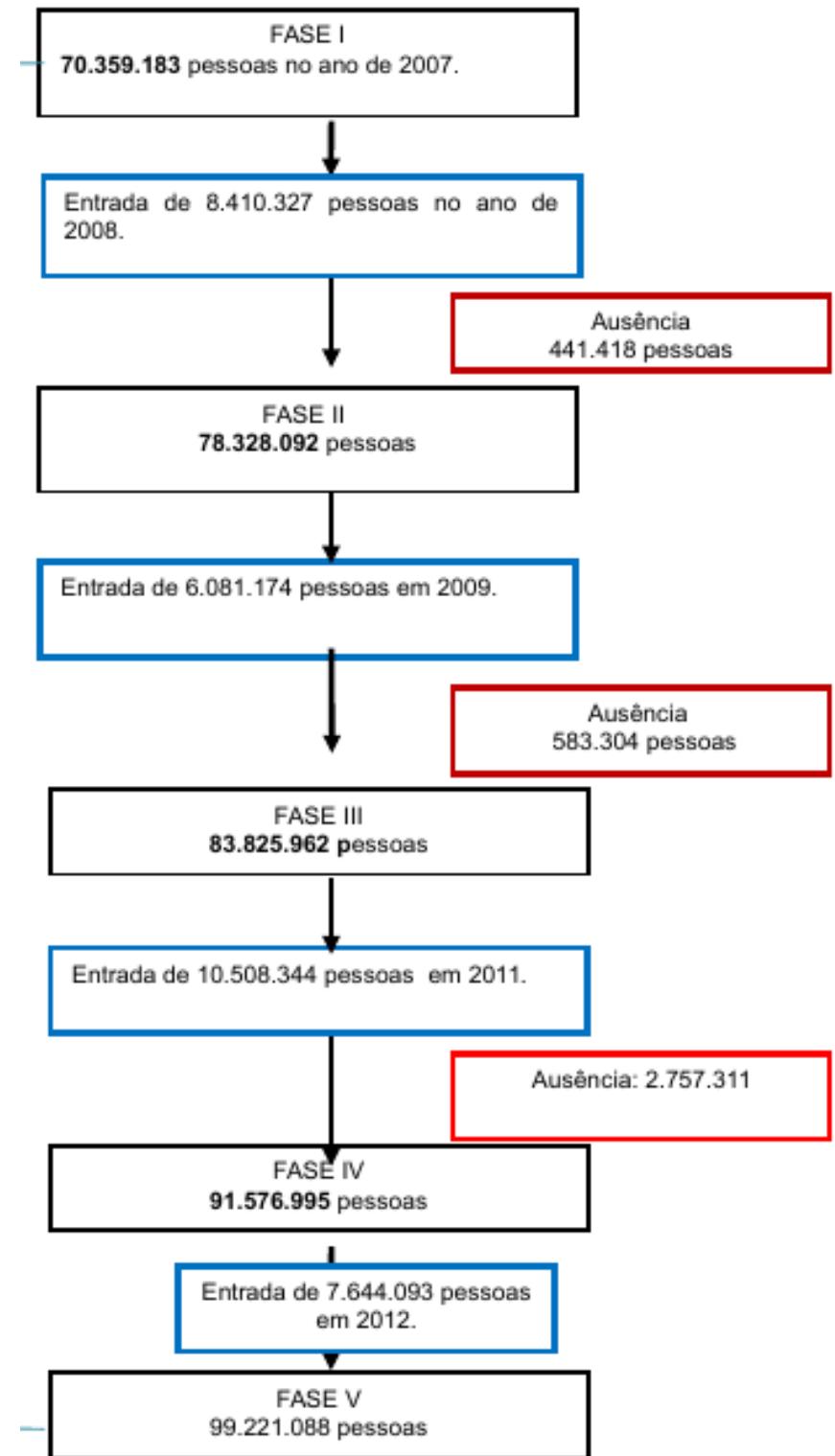
- People with at least one participation (basis 2012)



Cohort – in / out flow

Estimative (including 2010):
103,003,121 person

- Entry
- Exit



Some open problems...

- How to deal with family 'ramifications'?
- Analyses based on the whole family or only individuals?
- How to deal with missing attributes (especially NIS)?
- How to efficiently assess the quality of string attributes?
- How to increase accuracy when linking 'unknown' databases?

Thank you!
(Obrigado!)

marcoseb@dcc.ufba.br