THE ROYAL SOCIETY

LARGE-SCALE DATA FUSION OF BRAZILIAN SOCIOECONOMIC AND PUBLIC HEALTH DATABASES

Dr. Marcos Barreto

Newton International Fellow 2016-2018
Farr Institute of Health Informatics Research, University College London

Assistant Professor, Computer Science Department
Federal University of Bahia (UFBA), Salvador, Brazil

The Royal Society Meeting of Minds Conference
London, 31 May 2017

Ministério da Saúde
FIOCRUZ
Fundação Oswaldo Cruz

UFBA

UnB

Farr

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

SENAI CIMATEC SISTEMA FIEB
Federação das Indústrias do Estado da Bahia

SECRETARIA DE CIÊNCIA, TECNOLOGIA E INOVAÇÃO
BAHIA GOVERNO DO ESTADO

RIO+ WORLD CENTRE FOR SUSTAINABLE DEVELOPMENT

UNDP
Empowered lives. Resilient nations.

BILL & MELINDA GATES foundation

SUS Ministério da Saúde

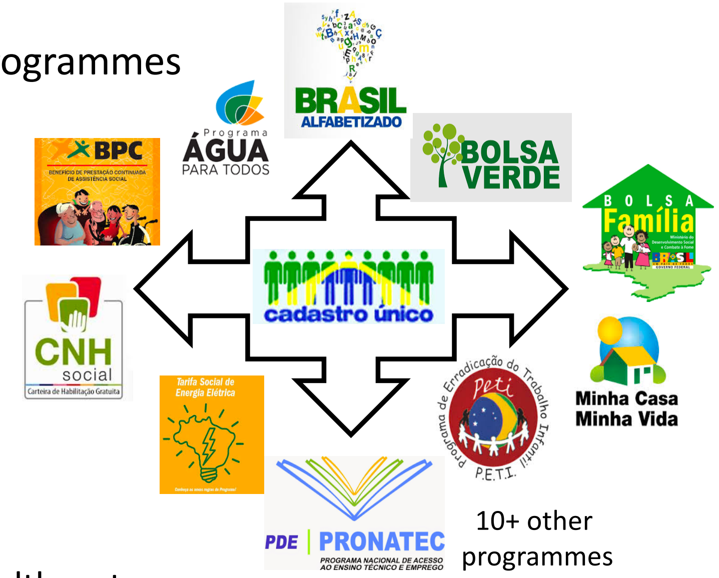
Ministério do Desenvolvimento Social e Combate à Fome

Ministério da Ciência e Tecnologia

GOVERNO FEDERAL BRASIL PÁTRIA EDUCADORA

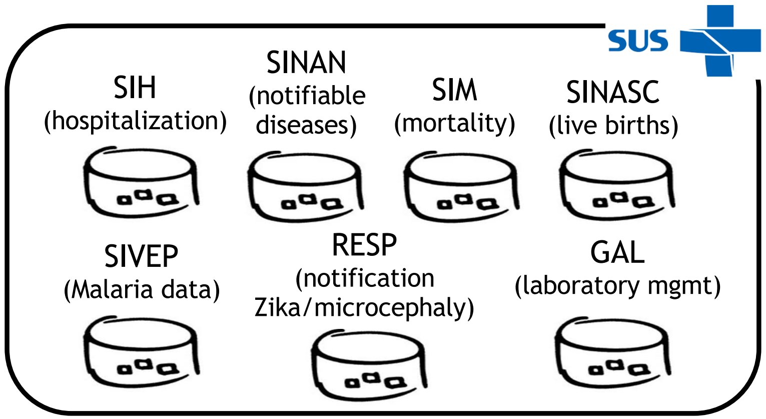
Brazilian governmental programmes & data

Social programmes



10+ other programmes

Public health system



SIH (hospitalization)

SINAN (notifiable diseases)

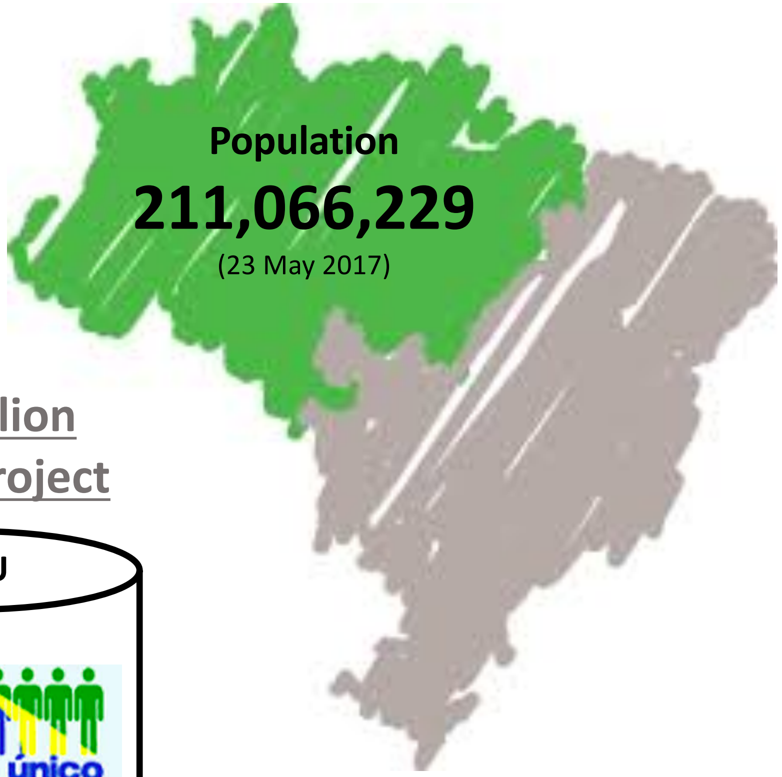
SIM (mortality)

SINASC (live births)

SIVEP (Malaria data)

RESP (notification Zika/microcephaly)

GAL (laboratory mgmt)



Population

211,066,229

(23 May 2017)

100 Million Cohort Project

CADU

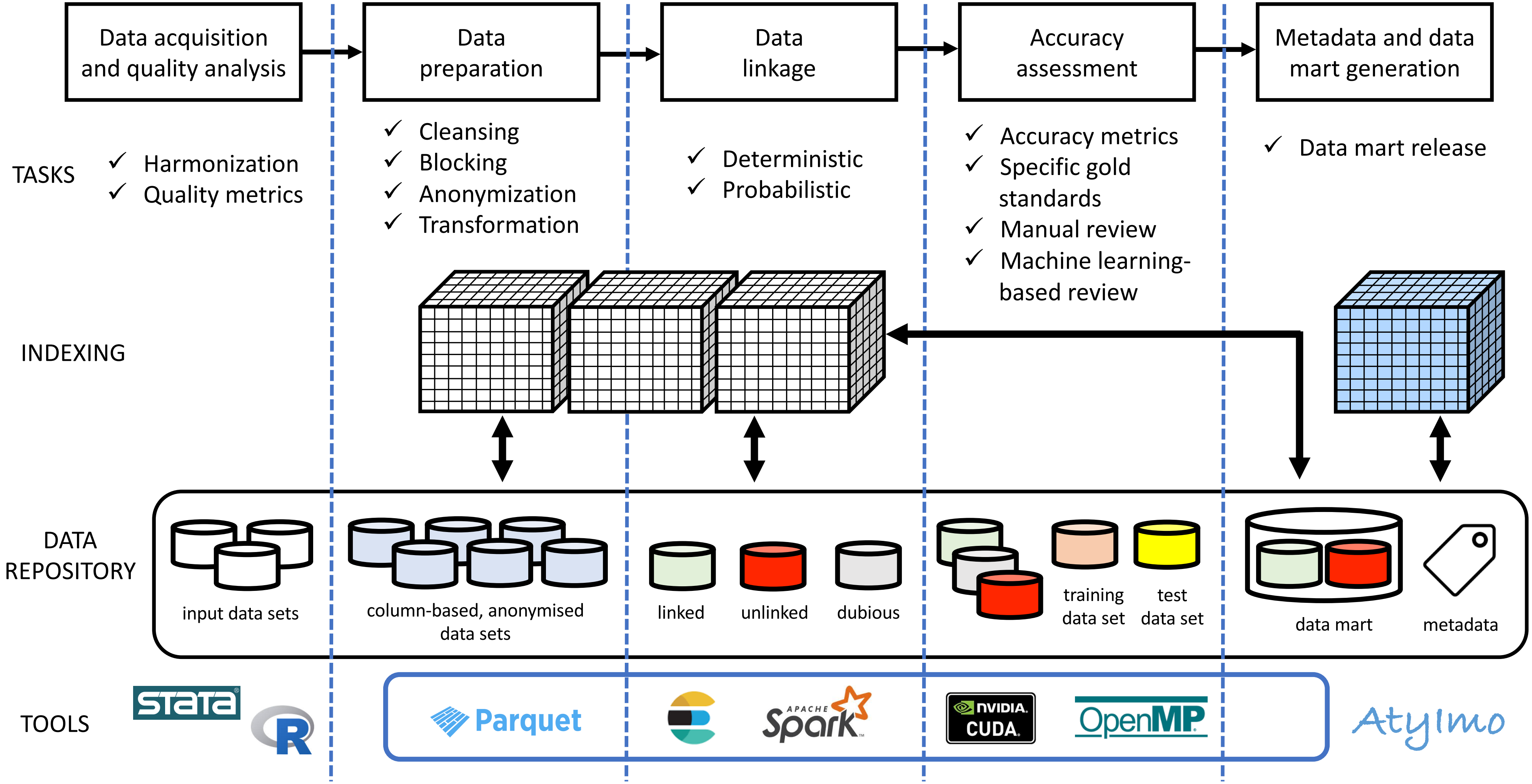
PBF

- ✓ Individuals registered in CADU
- ✓ Payments from Bolsa Família (conditional cash transfers)
- ✓ Period: 2003 – 2015
- ✓ **114 million individuals**
- ✓ ≅ 5,000 variables / individual

Overall goals

- ✓Besides building the 100 million cohort ...
- ✓Design an integration model comprising the cohort and governmental data.
- ✓Develop (probabilistic) data linkage tools.
- ✓Generate high accurate "data marts" (domain-specific data).
  - ✓Taking into account the absence of gold standards!
- ✓Provide support for other Brazil-UK ongoing projects:
  - ✓Long-term surveillance platform for Zika and microcephaly (live births 2001 – 2015, ≅ 80 million)
  - ✓Predictive analytics for Malaria eradication (2003 – 2016, ≅ 5,4 million cases)
  - ✓Population genomics and genetic epidemiology (EPIGEN-Brasil, 6,487 individuals)
  - ✓Integration and mining of Bioinformatics data (SAGASystem – Analysis of arbovirus genomes)
- ✓Design a reference platform to enable the operation of our data linkage centre (CIDACS).

Data linkage platform



Data acquisition and quality analysis → Data preparation → Data linkage → Accuracy assessment → Metadata and data mart generation

TASKS

- ✓ Harmonization
- ✓ Quality metrics

- ✓ Cleansing
- ✓ Blocking
- ✓ Anonymization
- ✓ Transformation

- ✓ Deterministic
- ✓ Probabilistic

- ✓ Accuracy metrics
- ✓ Specific gold standards
- ✓ Manual review
- ✓ Machine learning-based review

- ✓ Data mart release

INDEXING

DATA REPOSITORY

input data sets

column-based, anonymised data sets

linked

unlinked

dubious

training data set

test data set

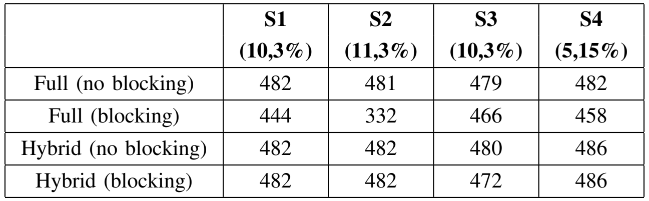
data mart

metadata

TOOLS

Accuracy assessment

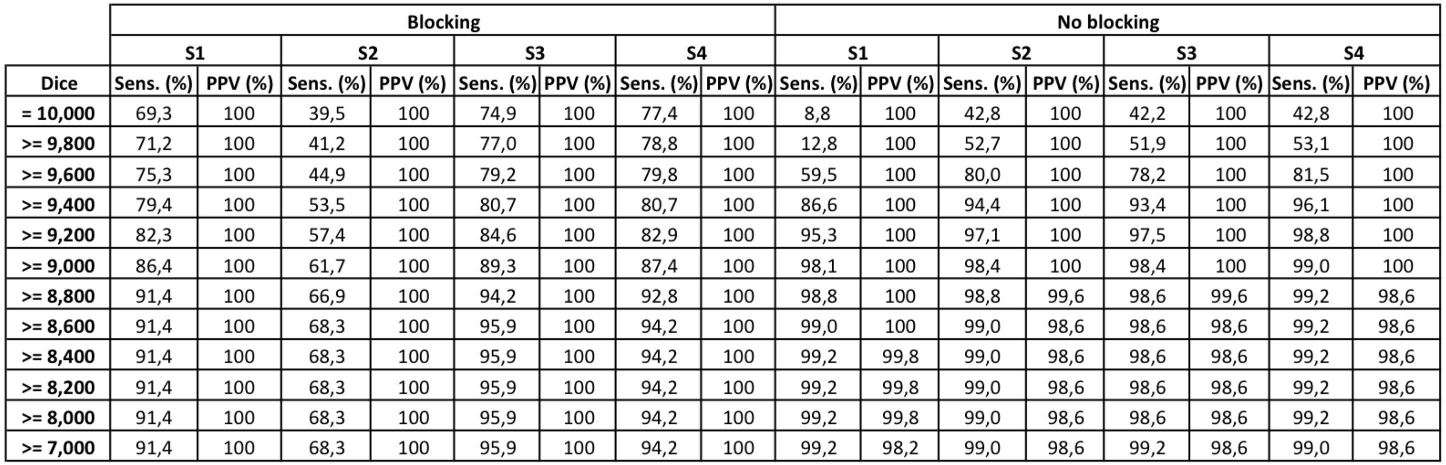
- ✓Metrics: sensitivity, specificity, positive predictive value (PPV), ROC curves.
- ✓Incremental samples from CADU cohort and public health databases.
- ✓Controlled (known) x uncontrolled scenarios



	S1 (10,3%)	S2 (11,3%)	S3 (10,3%)	S4 (5,15%)
Full (no blocking)	482	481	479	482
Full (blocking)	444	332	466	458
Hybrid (no blocking)	482	482	480	486
Hybrid (blocking)	482	482	472	486

Controlled scenario

- Rotavirus (486 positive exams + 200 random records)
- Hospitalizations (9,678 records)
- 4 simulation scenarios (*Si*)
- Different % of imputation errors
- Blocking X non-blocking linkage
- Goal: retrieve all the 486 records from hospitalization.



	Blocking								No blocking							
	S1		S2		S3		S4		S1		S2		S3		S4	
Dice	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**	**Sens. (%)**	**PPV (%)**
= 10,000	69,3	100	39,5	100	74,9	100	77,4	100	8,8	100	42,8	100	42,2	100	42,8	100
>= 9,800	71,2	100	41,2	100	77,0	100	78,8	100	12,8	100	52,7	100	51,9	100	53,1	100
>= 9,600	75,3	100	44,9	100	79,2	100	79,8	100	59,5	100	80,0	100	78,2	100	81,5	100
>= 9,400	79,4	100	53,5	100	80,7	100	80,7	100	86,6	100	94,4	100	93,4	100	96,1	100
>= 9,200	82,3	100	57,4	100	84,6	100	82,9	100	95,3	100	97,1	100	97,5	100	98,8	100
>= 9,000	86,4	100	61,7	100	89,3	100	87,4	100	98,1	100	98,4	100	98,4	100	99,0	100
>= 8,800	91,4	100	66,9	100	94,2	100	92,8	100	98,8	100	98,8	99,6	98,6	99,6	99,2	98,6
>= 8,600	91,4	100	68,3	100	95,9	100	94,2	100	99,0	100	99,0	98,6	98,6	98,6	99,2	98,6
>= 8,400	91,4	100	68,3	100	95,9	100	94,2	100	99,2	99,8	99,0	98,6	98,6	98,6	99,2	98,6
>= 8,200	91,4	100	68,3	100	95,9	100	94,2	100	99,2	99,8	99,0	98,6	98,6	98,6	99,2	98,6
>= 8,000	91,4	100	68,3	100	95,9	100	94,2	100	99,2	99,8	99,0	98,6	98,6	98,6	99,2	98,6
>= 7,000	91,4	100	68,3	100	95,9	100	94,2	100	99,2	98,2	99,0	98,6	99,2	98,6	99,0	98,6

TABLE II
LINKAGE FOR TUBERCULOSIS — UNCONTROLLED SCENARIO.

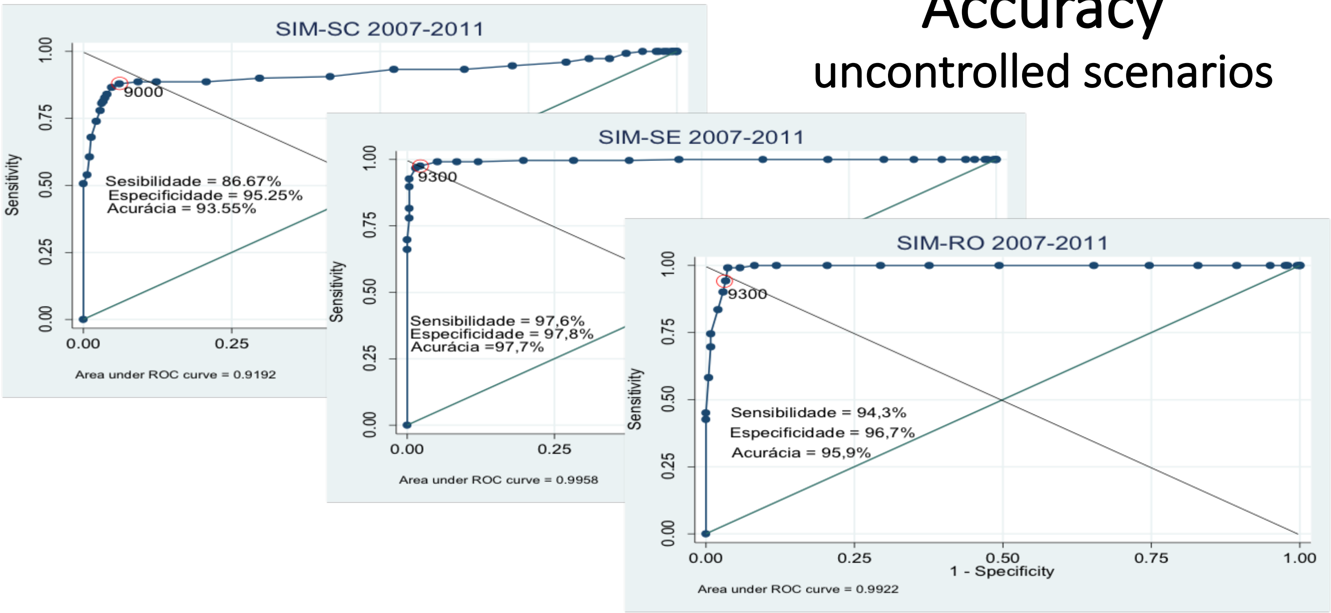
Databases (number of records)	Matched pairs		True positives (%)	
	Full	Hybrid	Full	Hybrid
CADU 2011 x SIH SE (1.447.512) x (49)	40	24	23 (57.5%)	23 (95.8%)
CADU 2011 x SIH SC (1.988.599) x (330)	140	95	83 (59.2%)	86 (90.5%)
CADU 2011 x SINAN SE (1.447.512) x (624)	398	311	309 (77.6%)	299 (96.1%)
CADU 2011 X SINAN SC (1.988.599) x (2.049)	661	500	551 (83.3%)	462 (92.4%)

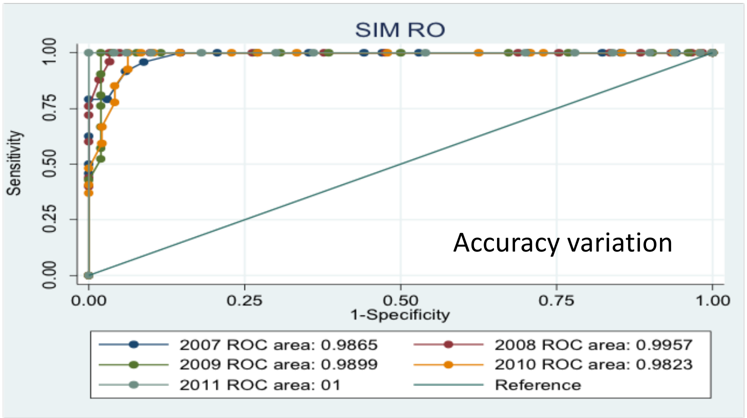
TABLE III
SUMMARY OF DICE COEFFICIENTS WITH BEST ACCURACY RESULTS.

States	SIH			SINAN		
	Dice	Sens.	PPV	Dice	Sens.	PPV
SE	9.400	95.6	95.0	9.300	96.7	95.9
SC	9.100	99.0	96.0	9.100	97.7	97.4
BA	9.100	98.5	97.9	9.200	95.7	95.5
RO	9.300	94.1	94.2	9.400	87.9	91.0

Accuracy

uncontrolled scenarios





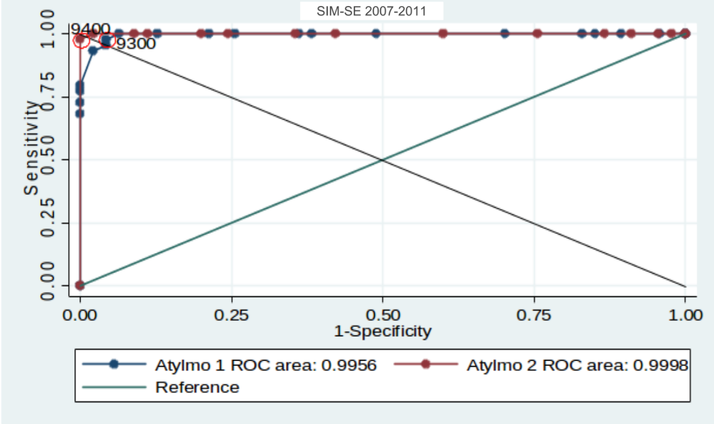
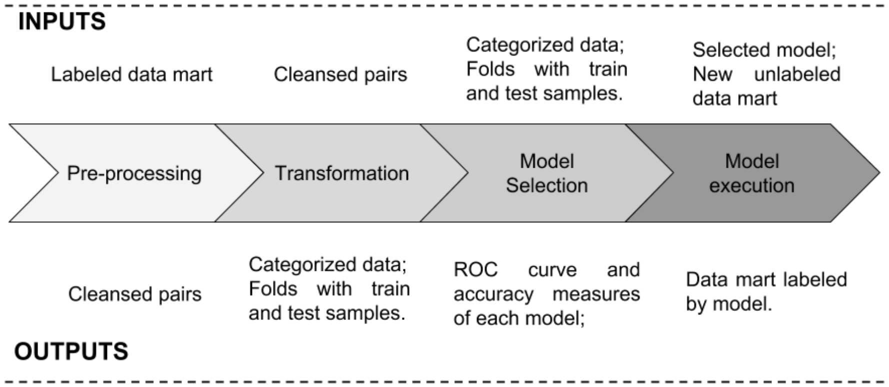


TABLE IV
COMPARATIVE ANALYSIS (ATYIMO V1 X ATYIMO V2).

States	AtyImo v1		AtyImo v2	
	Dice	ROC area	Dice	ROC area
SE	9.300	0.99	9.400	0.99
SC	9.000	0.96	9.100	0.99
RO	8.800	0.99	9.200	1

Current efforts

✓ Trainable model to accuracy assessment



INPUTS

Labeled data mart

Cleansed pairs

Categorized data; Folds with train and test samples.

Selected model; New unlabeled data mart

Pre-processing

Transformation

Model Selection

Model execution

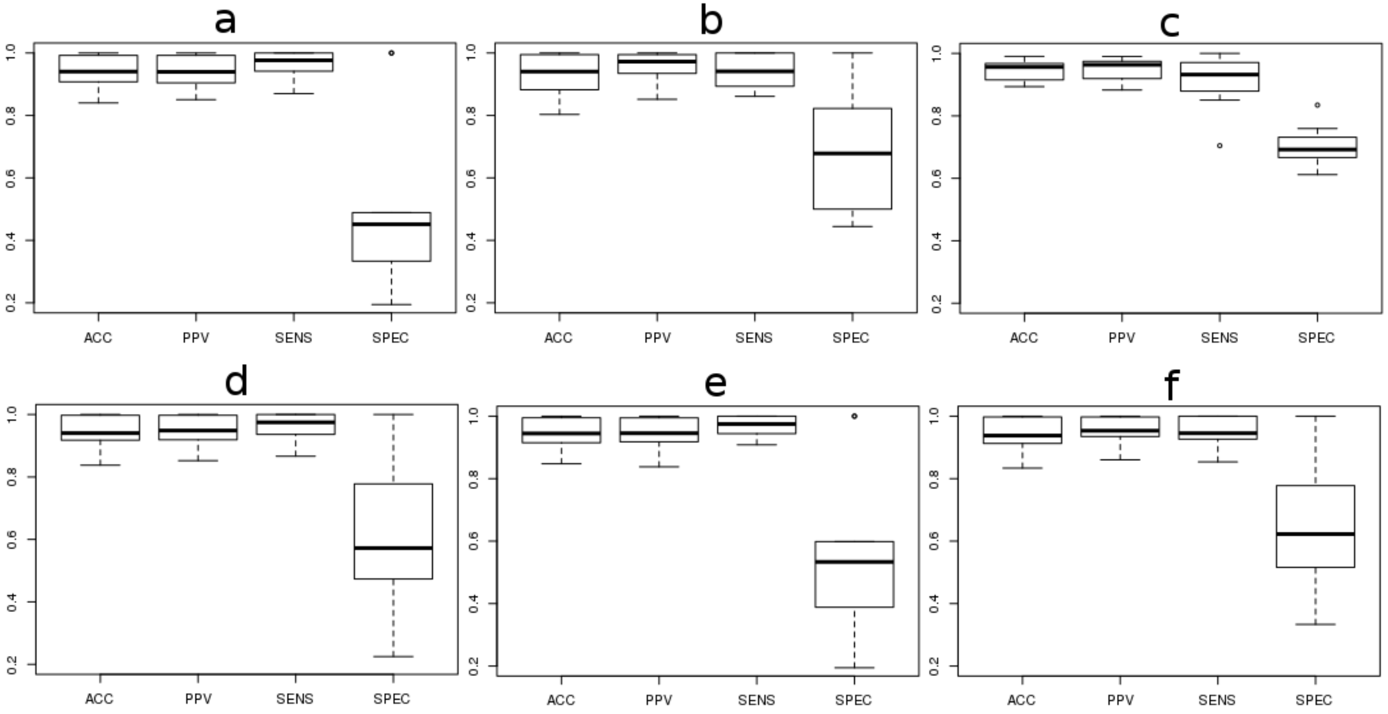
Cleansed pairs

Categorized data; Folds with train and test samples.

ROC curve and accuracy measures of each model;

Data mart labeled by model.

OUTPUTS



- 10-fold cross-validation
- 13,300 downto 7,880 records
- A – Decision trees
- B – Naïve Bayes
- C – Logistic regression
- D – Random forest
- E – Support vector machine
- F – Gradient boosted trees

Current efforts

- ✓Auto-tuning approach to ensure scalability over hybrid parallel architectures

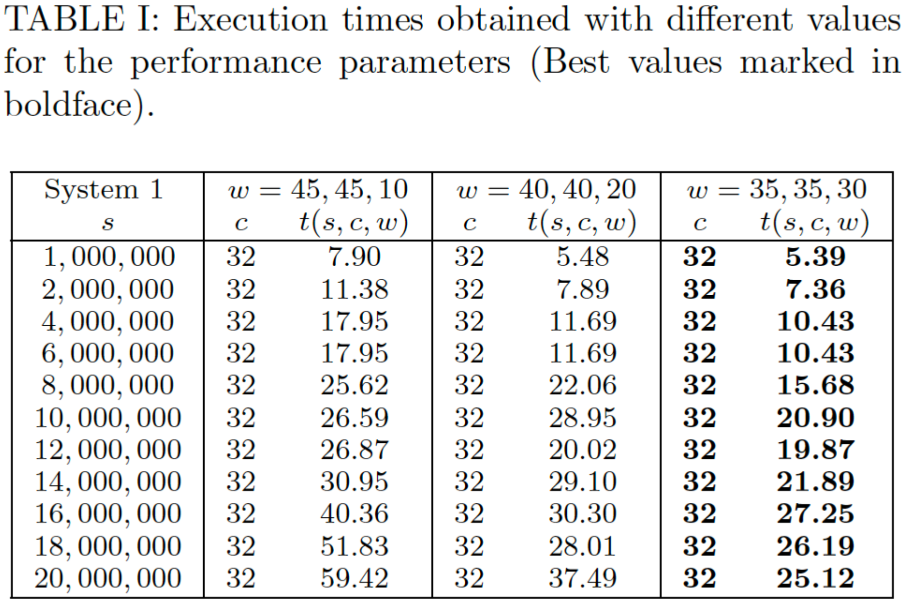
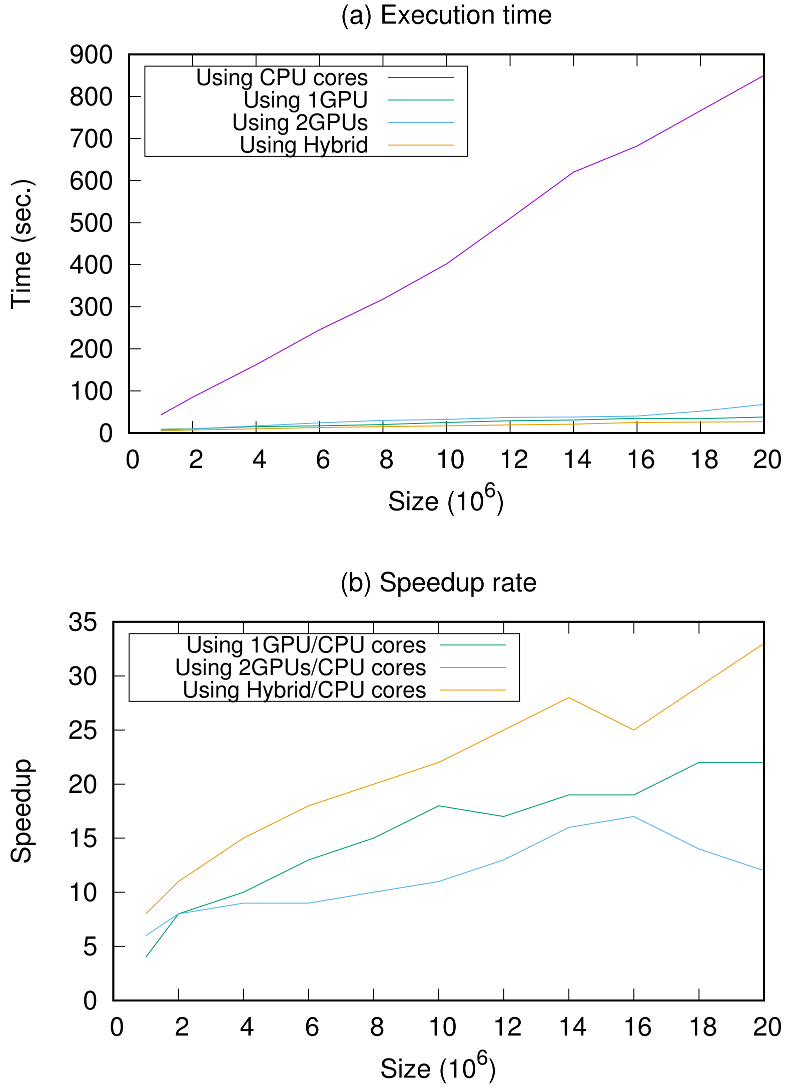


TABLE I: Execution times obtained with different values for the performance parameters (Best values marked in boldface).

System 1	$w = 45, 45, 10$		$w = 40, 40, 20$		$w = 35, 35, 30$	
s	c	$t(s, c, w)$	c	$t(s, c, w)$	c	$t(s, c, w)$
1,000,000	32	7.90	32	5.48	**32**	**5.39**
2,000,000	32	11.38	32	7.89	**32**	**7.36**
4,000,000	32	17.95	32	11.69	**32**	**10.43**
6,000,000	32	17.95	32	11.69	**32**	**10.43**
8,000,000	32	25.62	32	22.06	**32**	**15.68**
10,000,000	32	26.59	32	28.95	**32**	**20.90**
12,000,000	32	26.87	32	20.02	**32**	**19.87**
14,000,000	32	30.95	32	29.10	**32**	**21.89**
16,000,000	32	40.36	32	30.30	**32**	**27.25**
18,000,000	32	51.83	32	28.01	**32**	**26.19**
20,000,000	32	59.42	32	37.49	**32**	**25.12**



(a) Execution time

(b) Speedup rate

Getting access & contact

Marcos Barreto

m.barreto@ucl.ac.uk

Web: **www.dcc.ufba.br/~marcoseb**

cidacs
Centro de Integração de Dados
e Conhecimentos para a Saúde

http://www.cidacs.bahia.fiocruz.br/

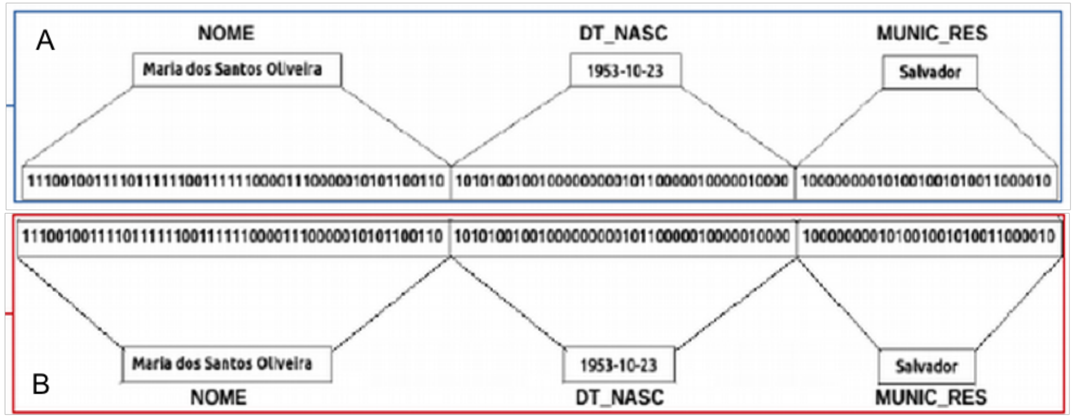
Public interface and Call for Projects to be released soon.

Prof. Mauricio Barreto

mauricio.barreto@bahia.fiocruz.br

Atyimo

- **Full probabilistic**: Sorensen (Dice) index applied to Bloom filters.



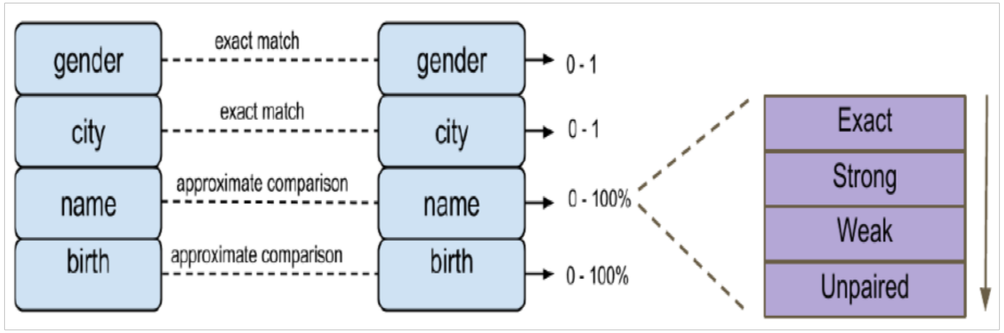
$$D_{a,b} = \frac{2h}{|a| + |b|} = [0, 1]$$

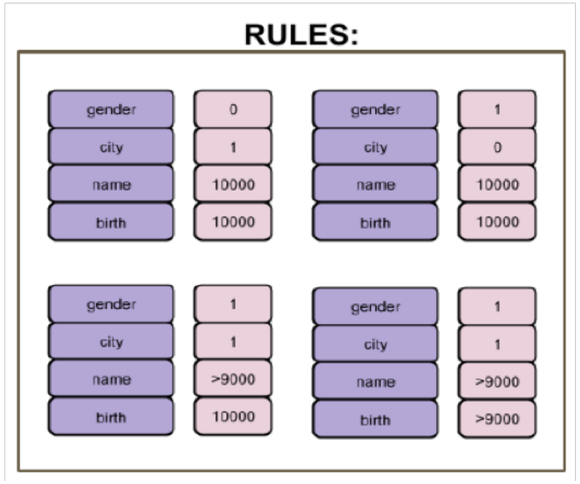
h = number of 1's at same position in both Bloom filters

a = number of 1's in Bloom filter A

b = number of 1's in Bloom filter B

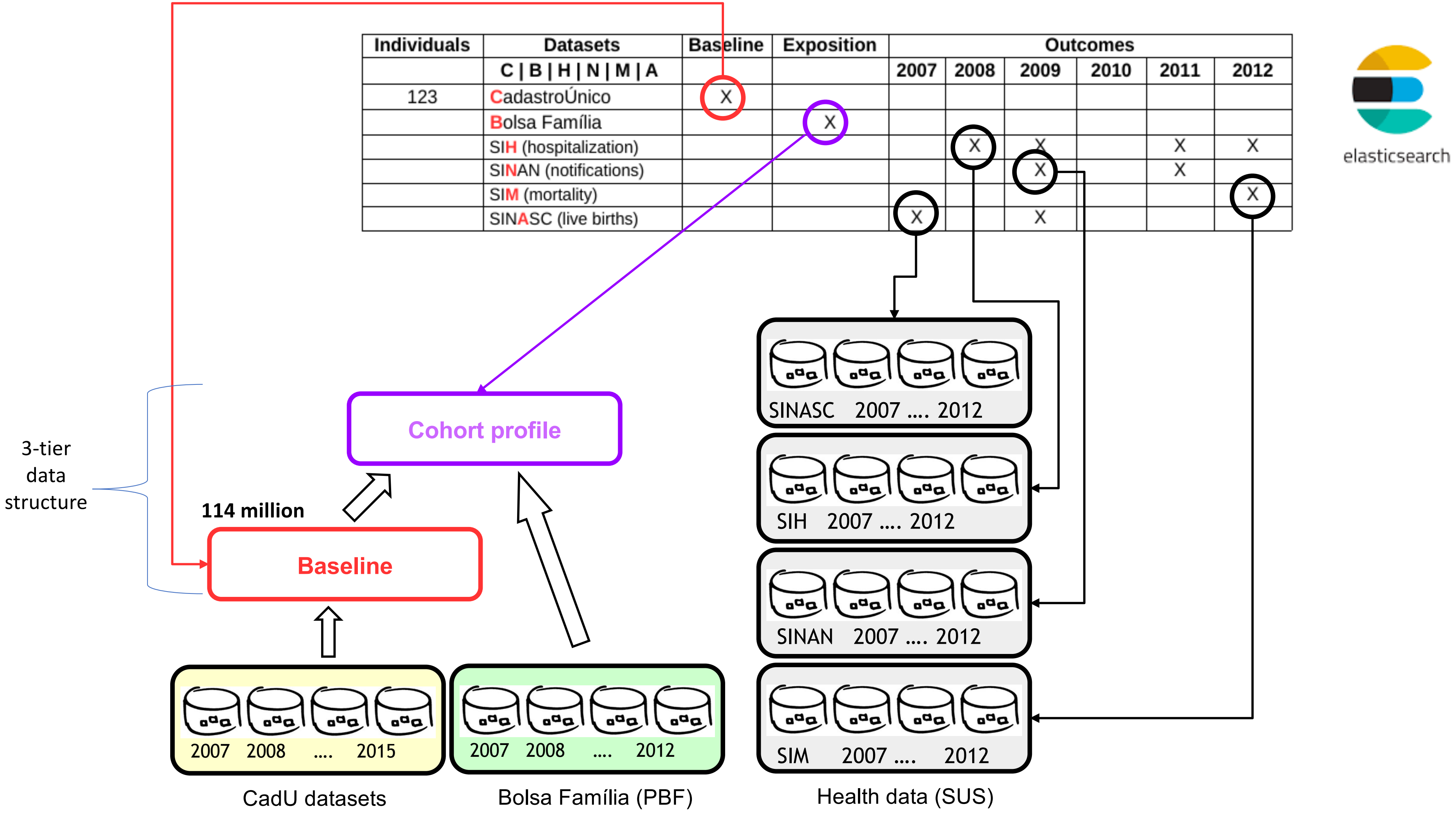
- **Hybrid approach**: individual comparison of attributes based on different rules




RULES:

gender	0	gender	1
city	1	city	0
name	10000	name	10000
birth	10000	birth	10000

gender	1	gender	1
city	1	city	1
name	>9000	name	>9000
birth	10000	birth	>9000



Individuals	Datasets	Baseline	Exposition	Outcomes					
	C | B | H | N | M | A			2007	2008	2009	2010	2011	2012
123	CadastroÚnico	X							
	Bolsa Família		X						
	SIH (hospitalization)				X	X		X	X
	SINAN (notifications)					X		X	
	SIM (mortality)								X
	SINASC (live births)			X		X			