Frontiers of Engineering for Development symposium:

Engineers as healthcare practitioners

WHAT TOOLS AND METHODS DO WE NEED TO ENABLE SECONDARY USE OF ROUTINE DATA?

Dr. Marcos Barreto

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

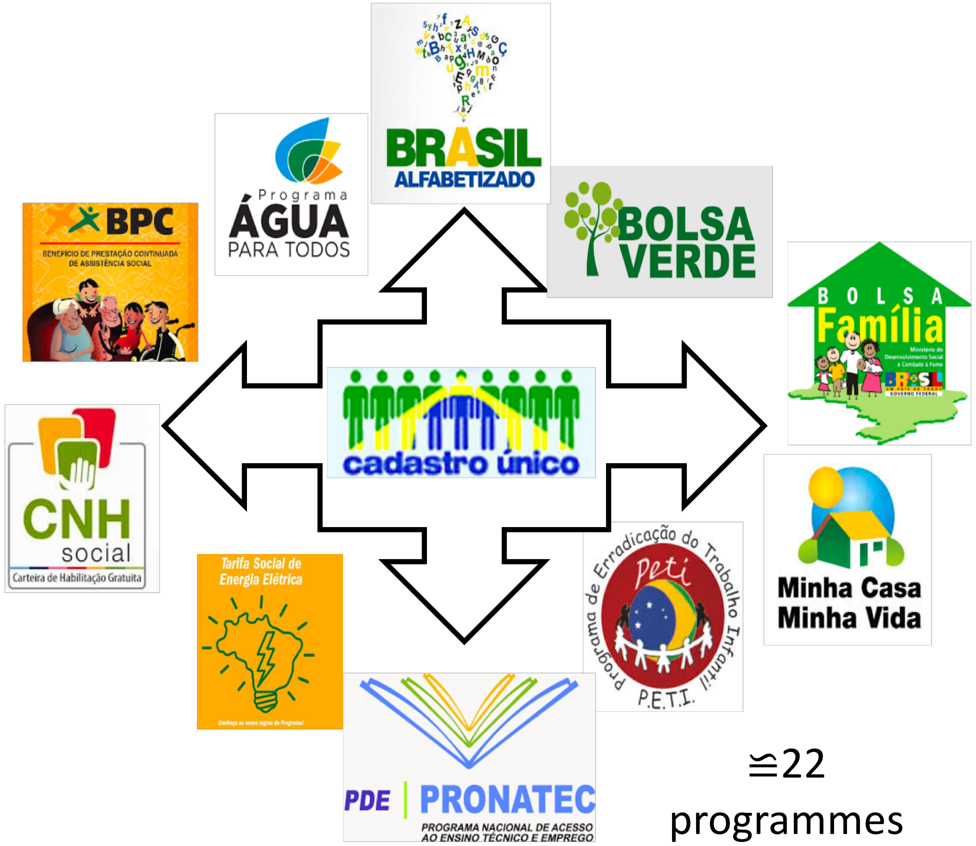
Farr Institute of Health Informatics Research, University College London
Newton International Fellow (The Royal Society, 2016-2018)

Ho Chi Minh City, Vietnam, 30 October to 2 November 2018

Socioeconomic and health care routine data in Brazil

Social programmes

- ✓ Targeted to poor and extremely poor families.
- ✓ Cadastro Único: central registrar for all programmes.


BRASIL ALFABETIZADO
Programa ÁGUA PARA TODOS
BPC
BENEFÍCIO DE PRESTAÇÃO CONTINUADA DE ASSISTÊNCIA SOCIAL
BOLSA VERDE
BOLSA Família
cadastro único
CNH social
Carteira de Habilitação Gratuita
Tarifa Social de Energia Elétrica
Programa de Erradicação do Trabalho Infantil
Peti
P.E.T.I.
Minha Casa Minha Vida
PDE | PRONATEC
PROGRAMA NACIONAL DE ACESSO AO ENSINO TÉCNICO E EMPREGO

≅22 programmes

Public health system (SUS)

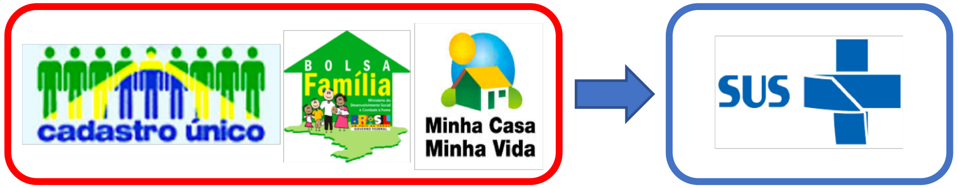
- ✓ Big and complex public health system.
  - from primary care to specialised transplantations.
- ✓ Used by approximately 77% of the Brazilian population (164 million people).

SUS

SIH (hospital episodes)
SINAN (notifiable diseases)
SIM (mortality)
SINASC (live births)
SISVAN (nutritional growth)
SIVEP (malaria notifications)
Vetores-malária (vector control)
RESP (microcephaly episodes -> Zika)
GAL (laboratory management)
…

Existing research platforms using these data

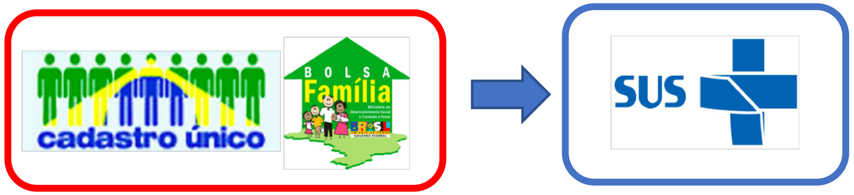
✓ The 100 Million Cohort



- ✓ Baseline: CadastroÚnico, 2001-2015, 114 million individuals x 367 atributes.
- ✓ Cohort: baseline + Bolsa Família (cash transfers) + Housing (MCMV), 2001 – 2015, 400 million records, 3,000 attributes.
- ✓ Used by +20 projects assessing the effects of social programmes on health outcomes.

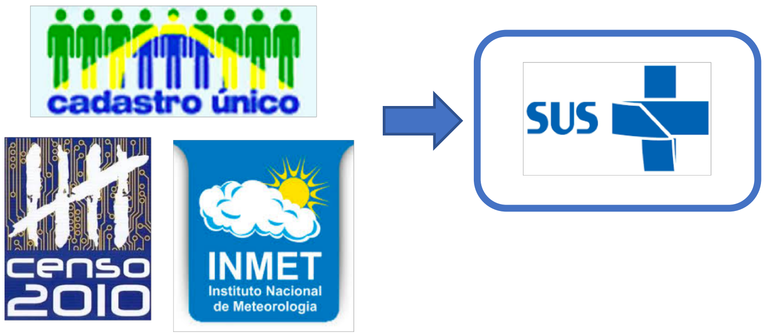
cidacs
Centro de Integração de Dados e Conhecimentos para a Saúde

✓ Zika surveillance (+ microcephaly)



- ✓ Birth cohort, 2001 – 2030, ≅80 million records.
- ✓ Morbidity, mortality, socioeconomic and service data.
- ✓ Focus on the triple epidemic (Zika, Dengue and Chikungunya) and health/educational outcomes related to microcephaly.

✓ Malaria linkage & analytics



- ✓ Malaria episodes (>5 million records) + mortality + socioeconomic + climate data, 2000 – 2018.
- ✓ Focus on i) data aggregation and ii) epidemic forecasting.

UFBA

FVS Fundação de Vigilância em Saúde

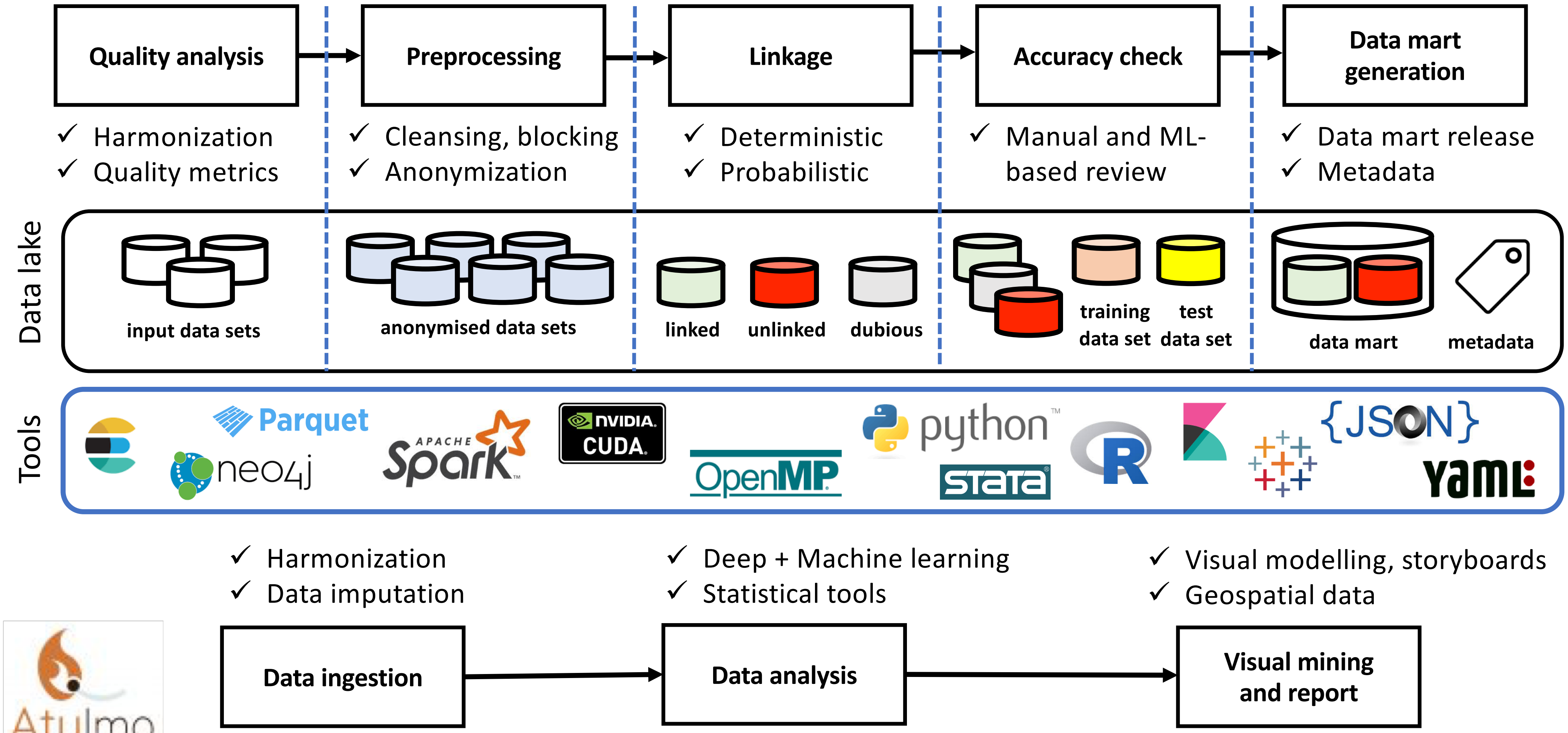
FIOCRUZ Fundação Oswaldo Cruz

BILL & MELINDA GATES foundation

UCL

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

Atylmo – Data linkage platform



Quality analysis
- ✓ Harmonization
- ✓ Quality metrics

Preprocessing
- ✓ Cleansing, blocking
- ✓ Anonymization

Linkage
- ✓ Deterministic
- ✓ Probabilistic

Accuracy check
- ✓ Manual and ML-based review

Data mart generation
- ✓ Data mart release
- ✓ Metadata

- ✓ Harmonization
- ✓ Data imputation

Data ingestion

- ✓ Deep + Machine learning
- ✓ Statistical tools

Data analysis

- ✓ Visual modelling, storyboards
- ✓ Geospatial data

Visual mining and report

Atylmo

<www.atyimolab.ufba.br>

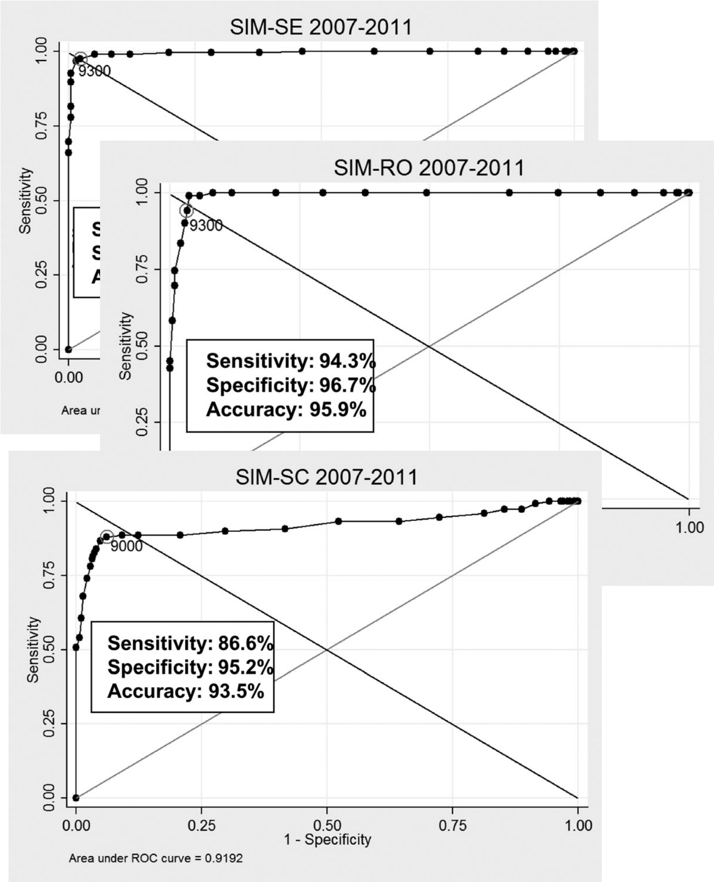
Data analytics pipeline

Example results

346 IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 22, NO. 2, MARCH 2018

On the Accuracy and Scalability of Probabilistic Data Linkage Over the Brazilian 114 Million Cohort

Robespierre Pita, Clícia Pinto, Samila Sena, Rosemeire Fiaccone, Leila Amorim, Sandra Reis, Mauricio L. Barreto, Spiros Denaxas, and Marcos Ennes Barreto



Sensitivity: 94.3%
Specificity: 96.7%
Accuracy: 95.9%

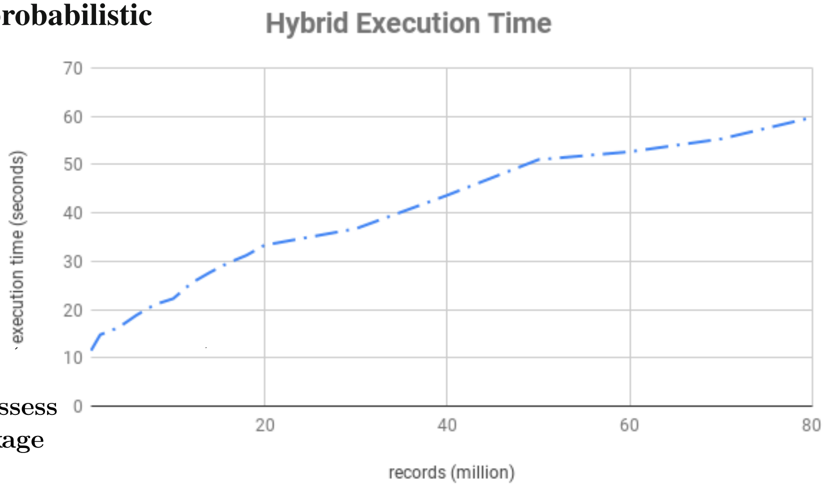
Sensitivity: 86.6%
Specificity: 95.2%
Accuracy: 93.5%

Area under ROC curve = 0.9192

Exploring hybrid parallel systems for probabilistic record linkage

Murilo Boratto[1] · Pedro Alonso[2] · Clicia Pinto[3] · Pedro Melo[3] · Marcos Barreto[3] · Spiros Denaxas[4]

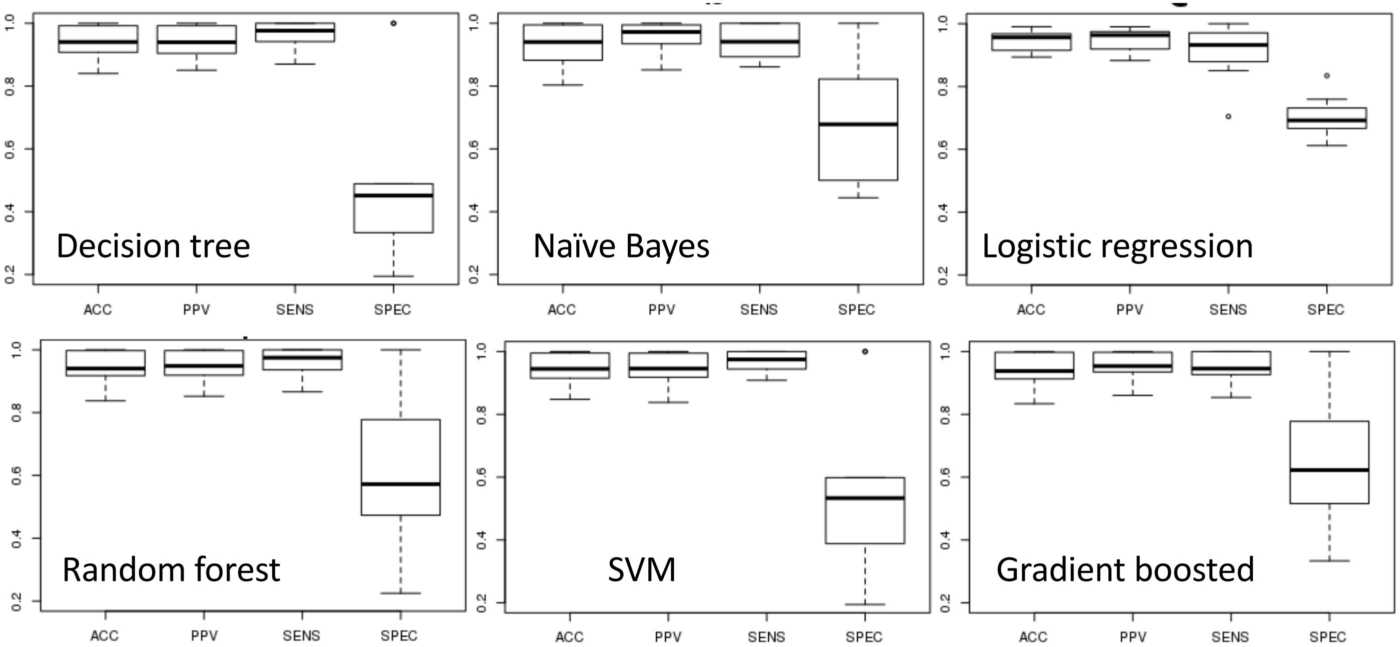
J Supercomput
https://doi.org/10.1007/s11227-018-2328-3



DOI: 10.1007/978-3-319-64283-3_16

A Machine Learning Trainable Model to Assess the Accuracy of Probabilistic Record Linkage

Robespierre Pita[1](✉), Everton Mendonça[1], Sandra Reis[2], Marcos Barreto[1,3], and Spiros Denaxas[3]



Biggest challenges

✓ Technical:

i) Large volumes of data hinder validation.

ii) Data heterogeneity hinders the adoption of generalizable machine learning methods.

✓ Behavioural:

✓ How to conciliate domain expertise (human) and powerful (but black-boxed) methods (machine) into accessible and effective analytics platforms to support routine research and policy-making over linked health care data?

Thank You.

Contact:
marcosb@ufba.br
www.atyimolab.ufba.br