DATA LINKAGE AND ANALYTICS EFFORTS OVER BRAZILIAN GOVERNMENTAL DATABASES

Prof. Dr. Marcos E. Barreto

AtyImoLab / LaSiD
Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade Federal da Bahia (UFBA)

Short bio

- ✓ BSc in Informatics (ULBRA, 1997), *distributed databases*.
- ✓ MSc and PhD in Computer Science (UFRGS, 2000, 2010), *high-performance computing*.
- ✓ Internship at Laboratoire d'Informatique de Grenoble, France (2000).
- ✓ Visiting professor at Ramon Llull University (Barcelona, 2002), *e-Science applications*.
- ✓ Associate Professor at UFBA (2010 – present), *data science*.
- ✓ Post-doctoral researcher at Institute of Health Informatics (University College London, 2016 – 2018), Newton International Fellow, *data linkage, machine learning*.
- ✓ PG Diploma on Data Science for Research in Health and Biomedicine (UCL, 2017-2018).

AtyImo

www.atyimolab.ufba.br

Salvador, Bahia

Why AtyImo

- ✓ Aty [tupi]: Unity
- ✓ Imo [iorubá]: Knowledge, Consciousness
- ✓ Átimo [portuguese]: suddenly, moment, a very brief period of time

Name coined by Robespierre Pita (2015)

Who we are


< BACK
Able Seamen
Sailmaker
Carpenter
Cook
1st Mate
2nd Mate
3rd Mate
Ordinary Seamen
Master
Apprentice
Steward

Who we are + What we do

PhD students

Robespierre Pita
Categorical data clustering

Clícia Pinto
Parallel data linkage

Elizabeth Gomes
High-performance plasma Physics

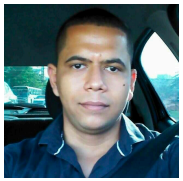
Everton Mendonça
Deep learning over EHR

Julio Oliveira
Integrative epidemiology

Master students

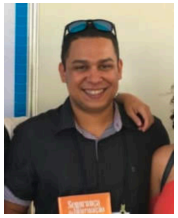
Juracy Bertoldo
Malaria forecasting models

Alberto Sironi
Visual data mining



Antonio Batista Jonatas Araujo
Cloud robotics

Danilo Azevedo
EHR over Blockchain



Patrick Ferraz
TBD

Undergraduate students

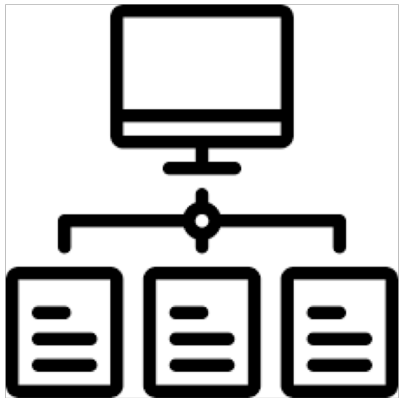
Paulo Costa
ML-based disaster warning system

João Gondim
Classification models for unlabelled data

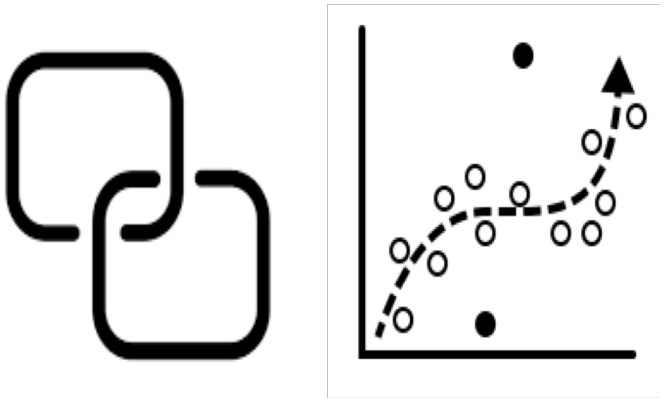
What we do

Atylmo

cloud
robotics



hybrid
parallel
computing



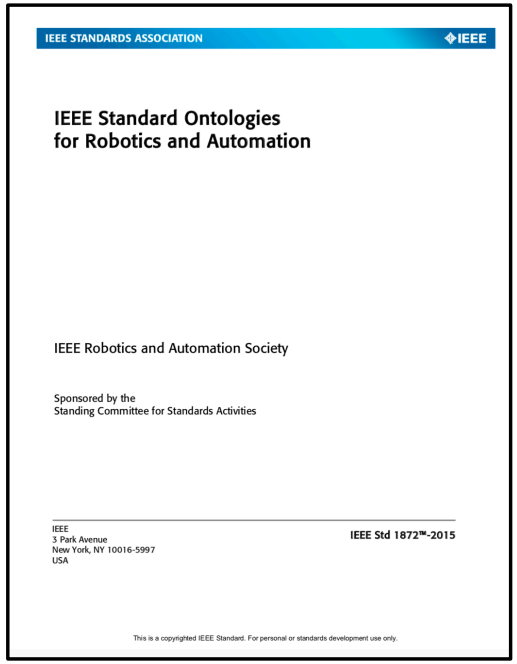
big data
linkage &
analytics

What we do

cloud
robotics

cloud robotics

- ✓ **IEEE Working Group on Ontologies for Robotics and Automation (ORA WG)**
- ✓ 2011 – 2015, +40 participants, Craig Schlenoff (NIST, USA), Edson Prestes (UFRGS, Brazil).
- ✓ Standard providing an overall ontology and associated methodology for knowledge representation and reasoning in R&A.


IEEE STANDARDS ASSOCIATION IEEE

IEEE Standard Ontologies for Robotics and Automation

IEEE Robotics and Automation Society

Sponsored by the Standing Committee for Standards Activities

IEEE
3 Park Avenue
New York, NY 10016-5997
USA

IEEE Std 1872™-2015

This is a copyrighted IEEE Standard. For personal or standards development use only.


IEEE Standards Association

Emerging Technology Award

presented to

IEEE Robotics and Automation Society (RAS) Ontologies for Robotics and Automation (ORA) Working Group

For developing the IEEE 1872-2015 Core Ontologies for Robotics and Automation standard, which is the first RAS standard and one of the first ontology-based IEEE standards.

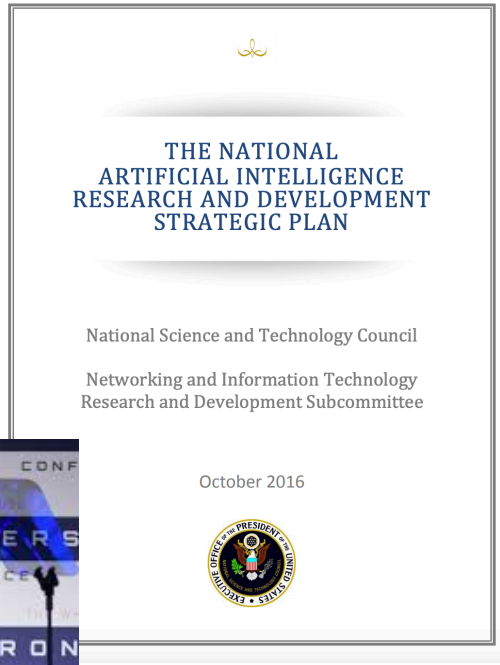
2015

IEEE

Bruce Kraemer
IEEE-SA President

Konstantinos Karachalios
IEEE-SA Managing Director

These guys are good!


THE NATIONAL ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT STRATEGIC PLAN

National Science and Technology Council

Networking and Information Technology Research and Development Subcommittee

October 2016

DOI: 10.1109/IEEESTD.2015.7084073

cloud robotics

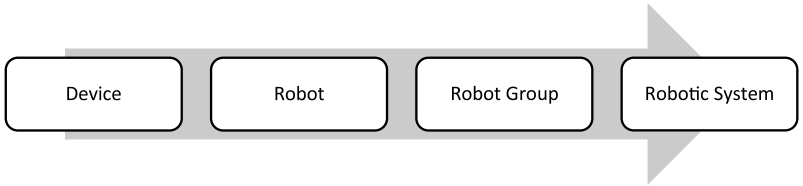


Fig. 2. Complexity axis and the main entities of the ontology.

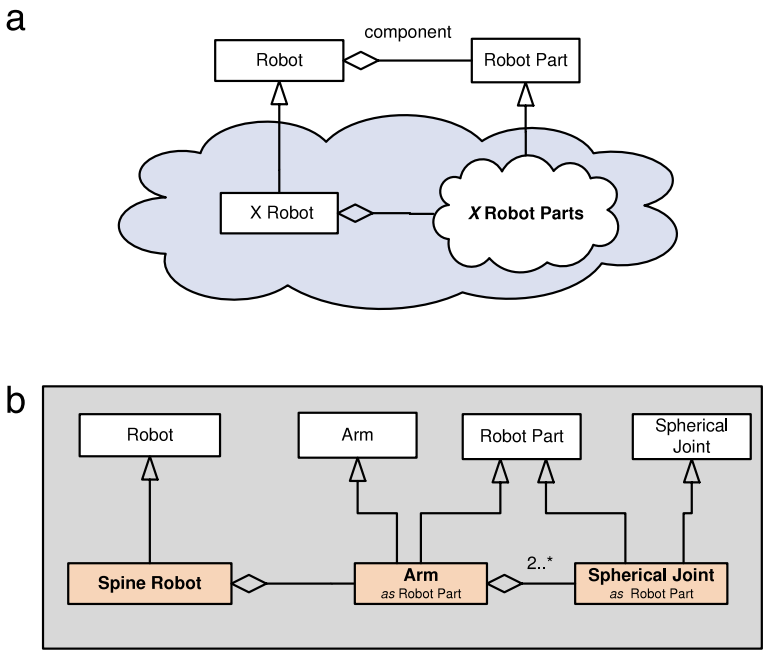


Fig. 4. Example of ontology pattern for specifying types of robots based on robot parts: (a) shows the general schema for extending the concept Robot with other kinds of robots based on robot parts, and (b) depicts pattern application example defining what is a spine robot.

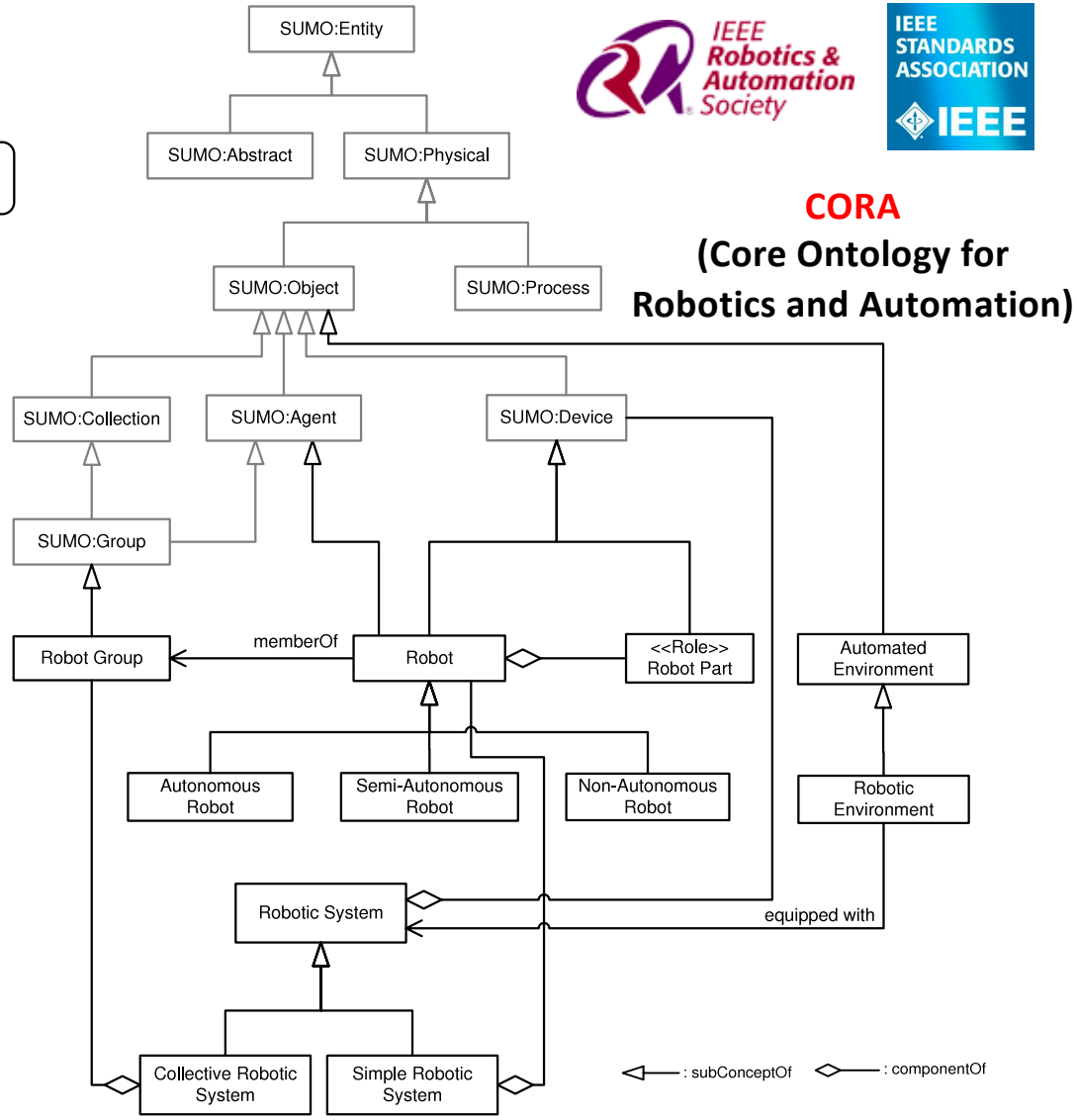


Fig. 6. Overview of the core ontology for robotics and automation. Concept imported from SUMO are prefixed to reflect that.

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Towards a core ontology for robotics and automation

Edson Prestes [a,*], Joel Luis Carbonera [a], Sandro Rama Fiorini [a], Vitor A. M. Jorge [a], Mara Abel [a], Raj Madhavan [b], Angela Locoro [c], Paulo Goncalves [d,e], Marcos E. Barreto [f], Maki Habib [g], Abdelghani Chibani [h], Sébastien Gérard [i], Yacine Amirat [h], Craig Schlenoff [j]

[a] Instituto de Informática, UFRGS, Brazil

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Ubiquitous robotics: Recent challenges and future trends

Abdelghani Chibani [a], Yacine Amirat [a], Samer Mohammed [a,*], Eric Matson [b,c,d], Norihiro Hagita [e], Marcos Barreto [f]

[a] University Paris-Est Créteil (UPEC), LISSI, Vitry-Sur-Seine, France

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Applied ontologies and standards for service robots

Tamás Haidegger [a,b,*], Marcos Barreto [c], Paulo Gonçalves [d], Maki K. Habib [e], Sampath Kumar Veera Ragavan [f], Howard Li [g], Alberto Vaccarella [h], Roberta Perrone [h], Edson Prestes [i]

[a] Óbuda University, Budapest, Hungary
[b] Austrian Center for Medical Innovation and Technology (ACMIT), Wiener Neustadt, Austria
[c] Distributed Systems Laboratory (LaSiD), UFBA, Brazil
[d] Polytechnic Institute of Castelo Branco / Technical University of Lisbon, Center of Intelligent Systems, IDMEC / LAETA, Portugal
[e] The American University in Cairo, Cairo, Egypt
[f] Monash University, Sunway campus, Selangor, Malaysia
[g] Department of Electrical and Computer Engineering, University of New Brunswick, Canada
[h] Department of Electronics, Information and Bioengineering (DEIB), NearLab Medical Robotics, Politecnico di Milano, Milan, Italy
[i] Instituto de Informática, UFRGS, Brazil

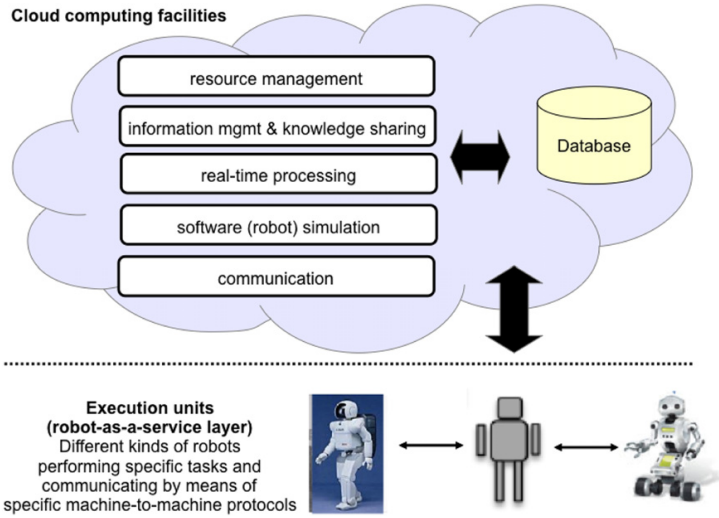
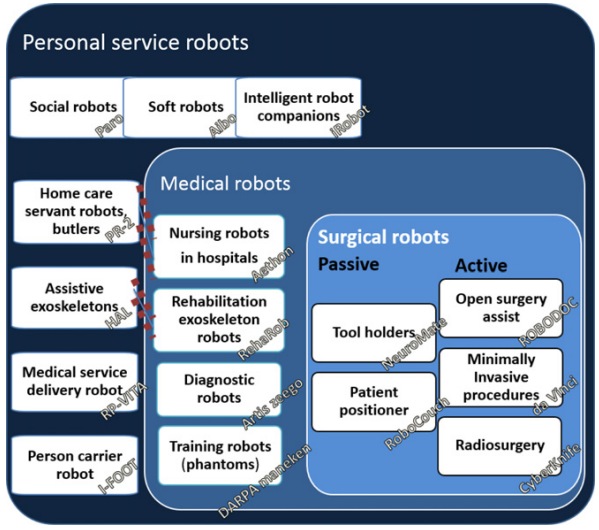
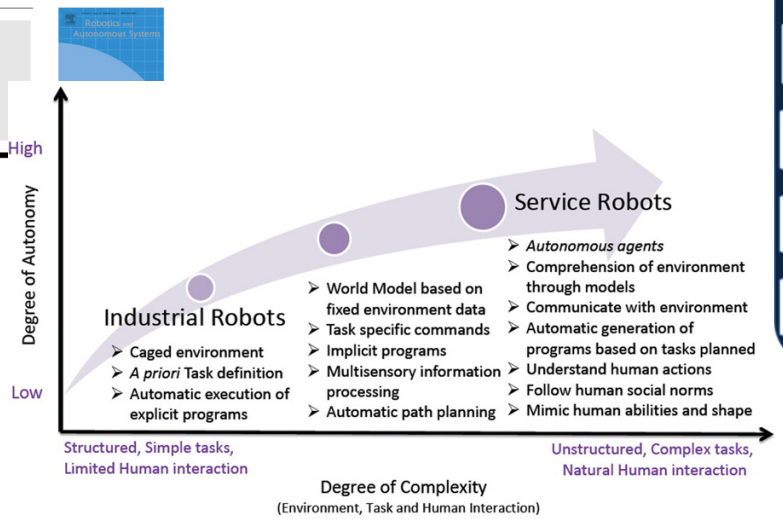


Fig. 1. A typical architecture for cloud robotics.





cloud robotics

An IEEE Standard Ontology for Robotics and Automation

Craig Schlenoff‡, Edson Prestes≀, Raj Madhavan‡, Paulo Goncalves†, Howard Li◇, Stephen Balakirsky‡, Thomas Kramer‡ and Emilio Migueláñez§

Nesting the Context for Pervasive Robotics

Tamás Haidegger*
Budapest Univ. of Technology and Economics, BME IIT
Budapest, Hungary
haidegger@ieee.org

Marcos E. Barreto
Computer Science Dept., Federal University of Bahia
UFBA, Salvador, Brazil
marcoseb@dcc.ufba.br

Paulo J.S. Gonçalves
Polytechnic Institute of Castelo Branco, TU Lisbon
IDMEC/LAETA, Lisbon, Portugal
paulo.goncalves@ipcb.pt

Maki Habib
The American University in Cairo
Egypt
maki@aucegypt.edu

Veera Ragavan
Monash University, Sunway campus
Selangor, Malaysia
veera.ragavan@monash.edu

Craig Schlenoff
Intelligent Systems Division, National Inst. of Standards and Technology
NIST, Gaithersburg, USA
craig.schlenoff@nist.gov

Alberto Vaccarella
Bioengineering Dept., Politecnico di Milano
Milan, Italy
vaccarella.alberto@gmail.com

Edson Prestes
Instituto de Informtica
UFRGS, Brazil
edson.prestes@ieee.org

Towards an Upper Ontology and Methodology for Robotics and Automation

Edson Prestes[1], Craig Schlenoff[2], Marcos Barreto[3], Abdelghani Chibani[4], Sébastien Gérard[5], Raj Madhavan[6], Alessandro Saffiotti[7], Ricardo Sanz[8], Yacine Amirat[4].

2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
November 3-7, 2013. Tokyo, Japan

Defining Positioning in a Core Ontology for Robotics*

Joel Luis Carbonera†, Sandro Rama Fiorini†, Edson Prestes†, Vitor A. M. Jorge†, Mara Abel†, Raj Madhavan‡, Angela Locoro§, Paulo Gonçalves¶, Tamás Haidegger‖, Marcos E. Barreto**, Craig Schlenoff††

Robot Ontologies for Sensor- and Image-Guided Surgery

Tamás Haidegger*, Marcos Barreto†, Paulo J. S. Gonçalves‡, Maki K. Habibℵ, S. Veera Ragavan£, Howard Li§, Alberto Vaccarella¶, Roberta Perrone¶, Edson Prestes▽

HUMANITARIAN TECHNOLOGY

2014 Humanitarian Robotics and Automation Technology Challenge

By Raj Madhavan, Lino Marques, Edson Prestes, Prithviraj Dasgupta, Gonçalo Cabrita, David Portugal, Bruno Gouveia, Vitor Jorge, Renan Maffei, Guilherme Franco, and Jose Garcia

Figure 1. The HRATC 2014 robot platform.

Figure 3. An HRATC framework main window.

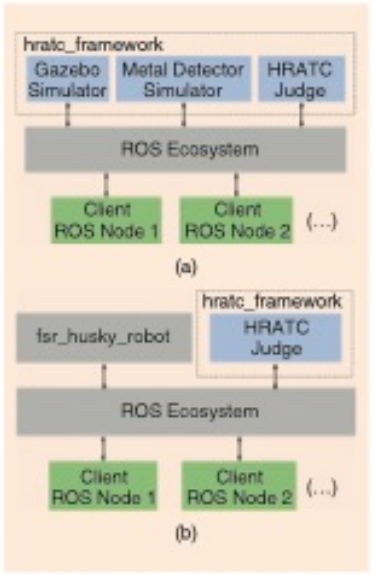


Figure 2. The software architecture for (a) the simulation and (b) the testing phase.

cloud robotics

- **IEEE Working Group on Autonomous Robotics**.
- **IEEE Working Group on Robot Task Representation**.
- **IEEE Working Group on Ethical Design and Application of Robotic Systems**.
- **IEEE Working Group on Personal Data Privacy Process**.

2017 26th IEEE International Symposium on
Robot and Human Interactive Communication (RO-MAN)
Lisbon, Portugal, Aug 28 - Sept 1, 2017.

Ontology for Autonomous Robotics

Joanna Isabelle Olszewska, Marcos Barreto, Julita Bermejo-Alonso, Joel Carbonera, Abdelghani Chibani, Sandro Fiorini, Paulo Goncalves, Maki Habib, Alaa Khamis, Alberto Olivares, Edison Pignaton de Freitas, Edson Prestes, S. Veera Ragavan, Signe Redfield, Ricardo Sanz, Bruce Spencer, and Howard Li

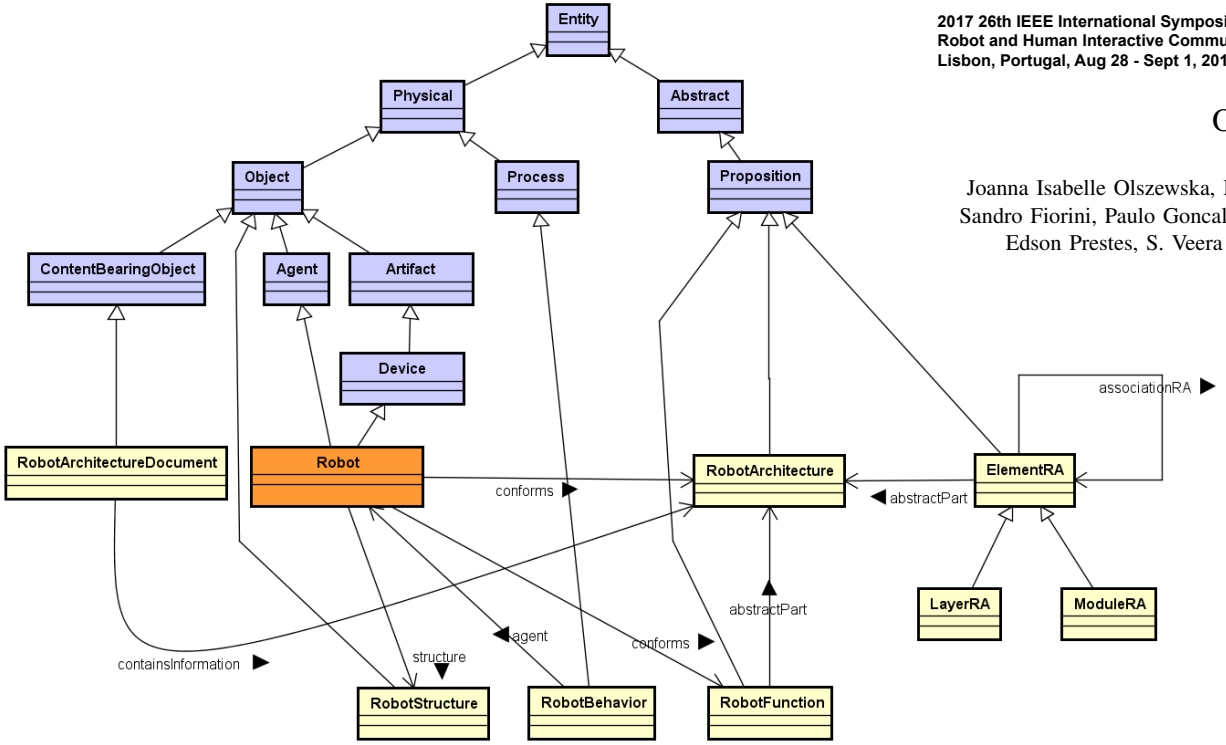


Fig. 1: Overview of the ontology's taxonomy and relations depicted in standard UML class diagram notation. Blue boxes are concepts from SUMO. Orange boxes are concepts from CORA. Yellow boxes are concepts from ROA Ontology. Almost all relations are imported from SUMO/CORA, with exception of *structure* and *associationRA*.

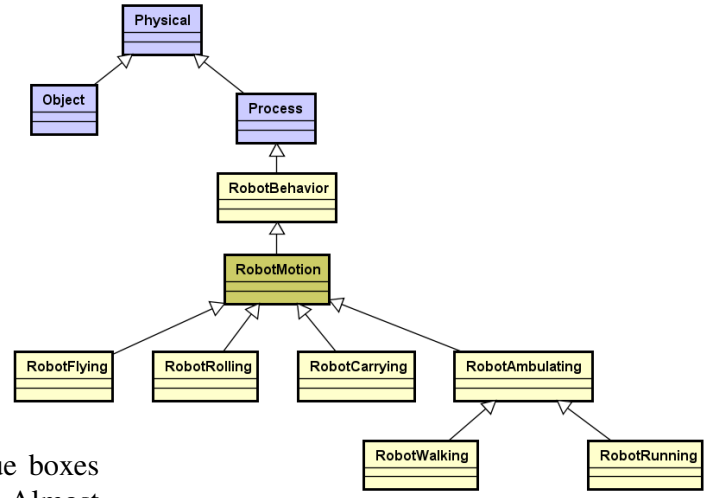


Fig. 2: Robot motion taxonomy.

cloud robotics

- **IEEE Working Group on Autonomous Robotics**.
- **IEEE Working Group on Robot Task Representation**.

An Ontology for Computational Robot Architectures

Sandro Rama Fiorini*,[1], Joel Luís Carbonera[2], S. Veera Ragavan[3], Joanna Isabelle Olszewska[4], Paulo Gonçalves[5], Abdelghani Chibani[1], Yacine Amirat[1], Marcos Barreto[6], Howard Li[7], and Edson Prestes[8]

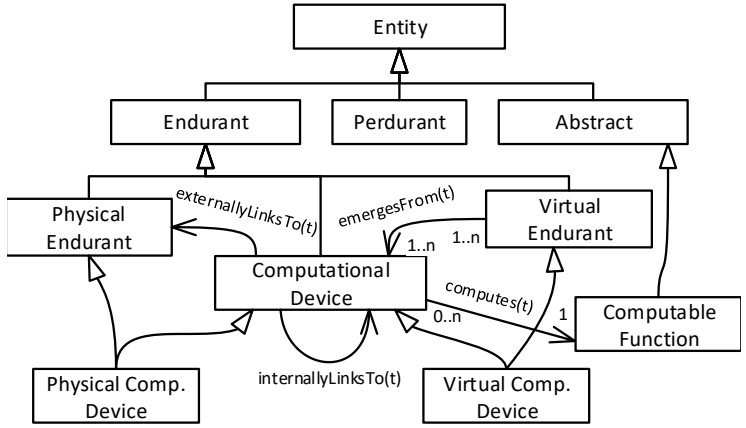


Fig. 2. Taxonomy of main entities related to ComputationalDevice.

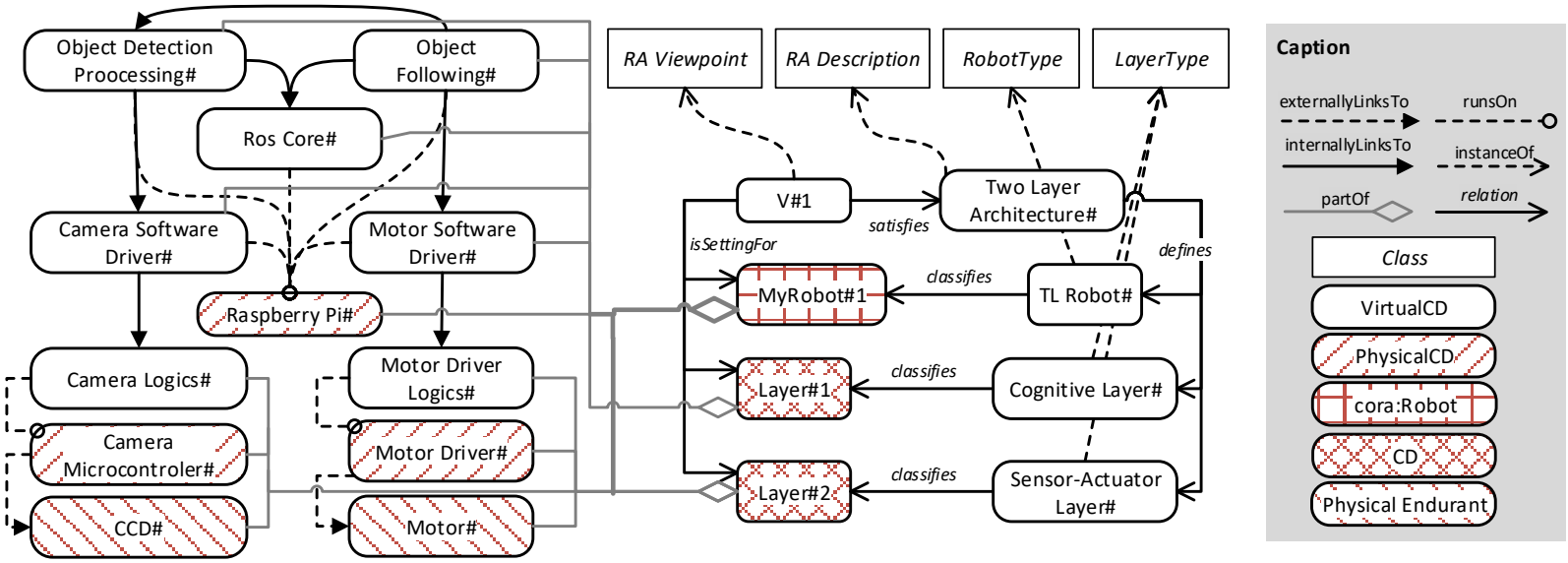
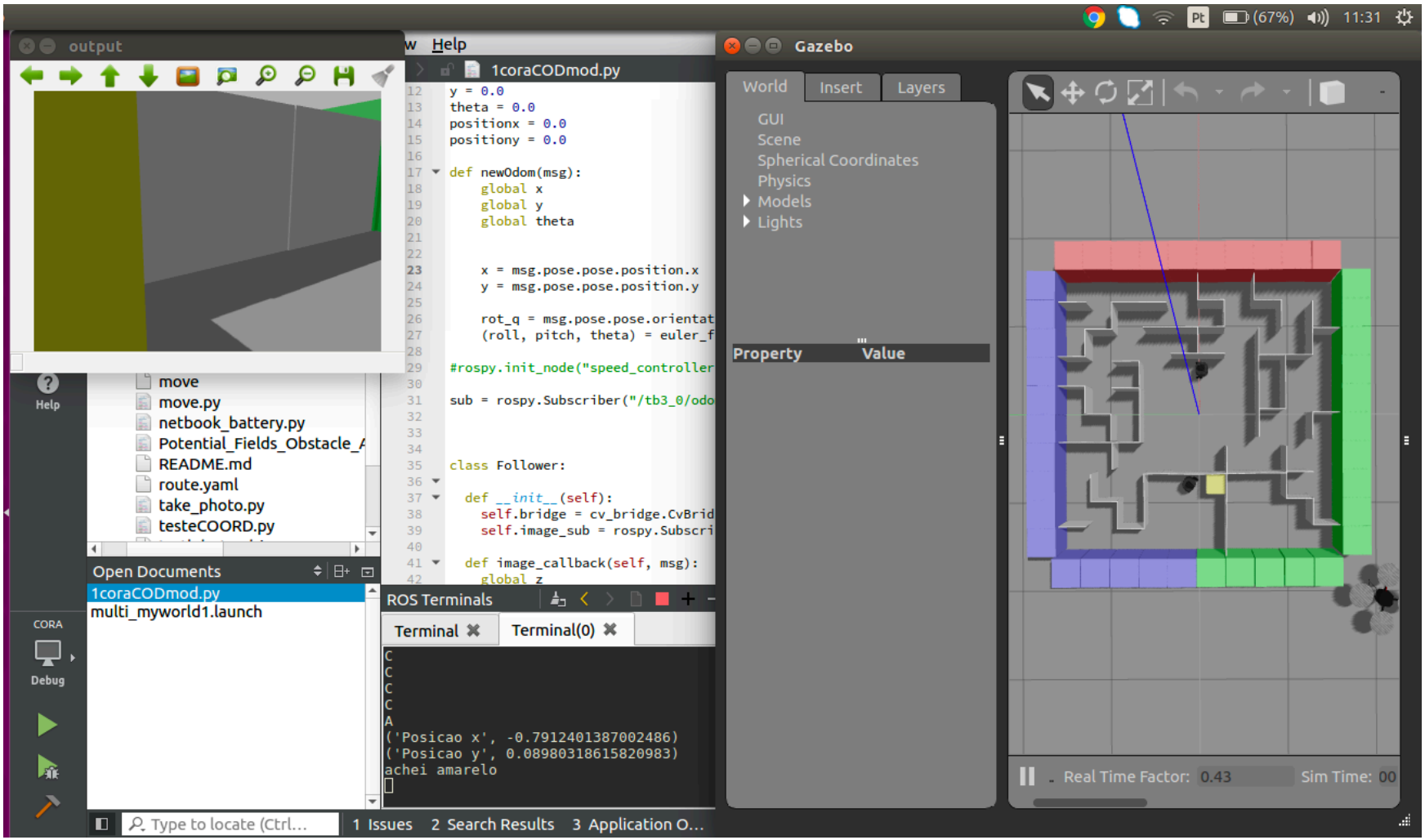


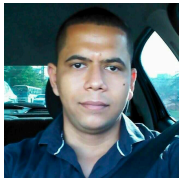
Fig. 3. Instantiation of MyRobot#1 and its description using OCRA. The caption is at the bottom. Rounded boxes are instances and rectangles are classes. The type of each individual is given by its background colour (i.e. pattern) or by the instanceOf relation.

cloud robotics

✓ design and validation of an autonomous robotics cloud-based architecture

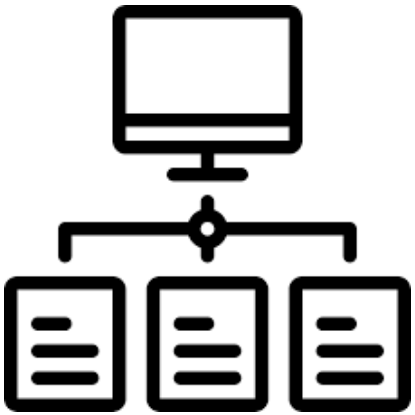


Antonio Batista



Jonatas Araujo

What we do



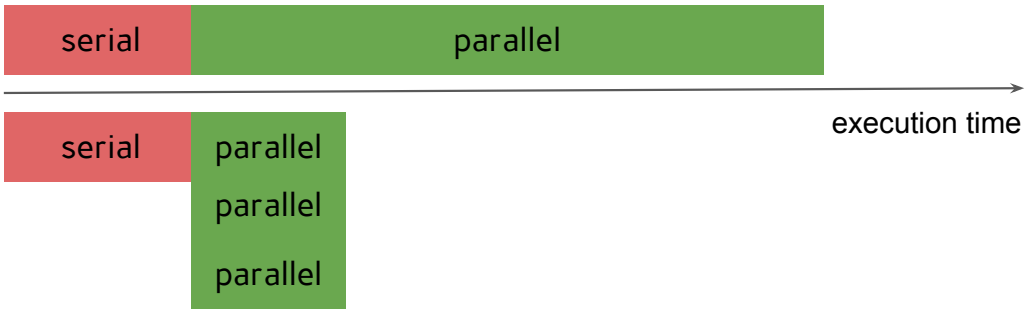
hybrid
parallel
computing

hybrid parallel computing

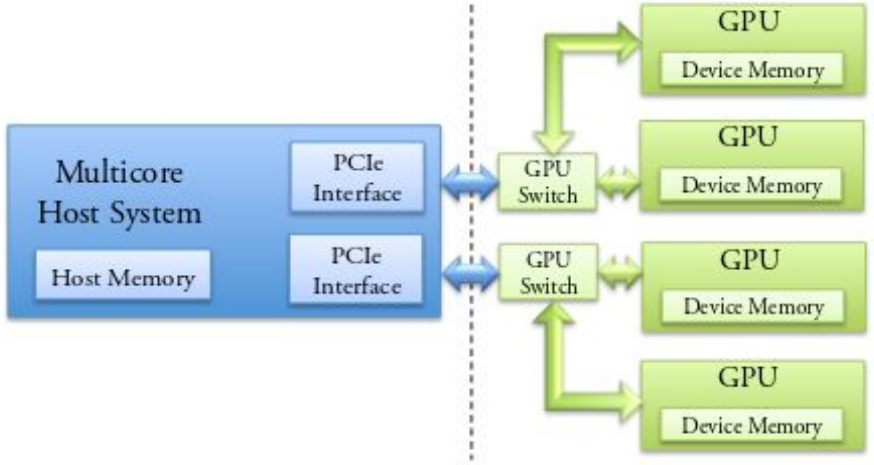
$$time = time_{serial} + \frac{time_{parallel}}{nproc}$$

✓ **How to efficiently explore hybrid architectures?**

$$Speedup \leq \frac{1}{\frac{p}{n} + (1 - p)}$$



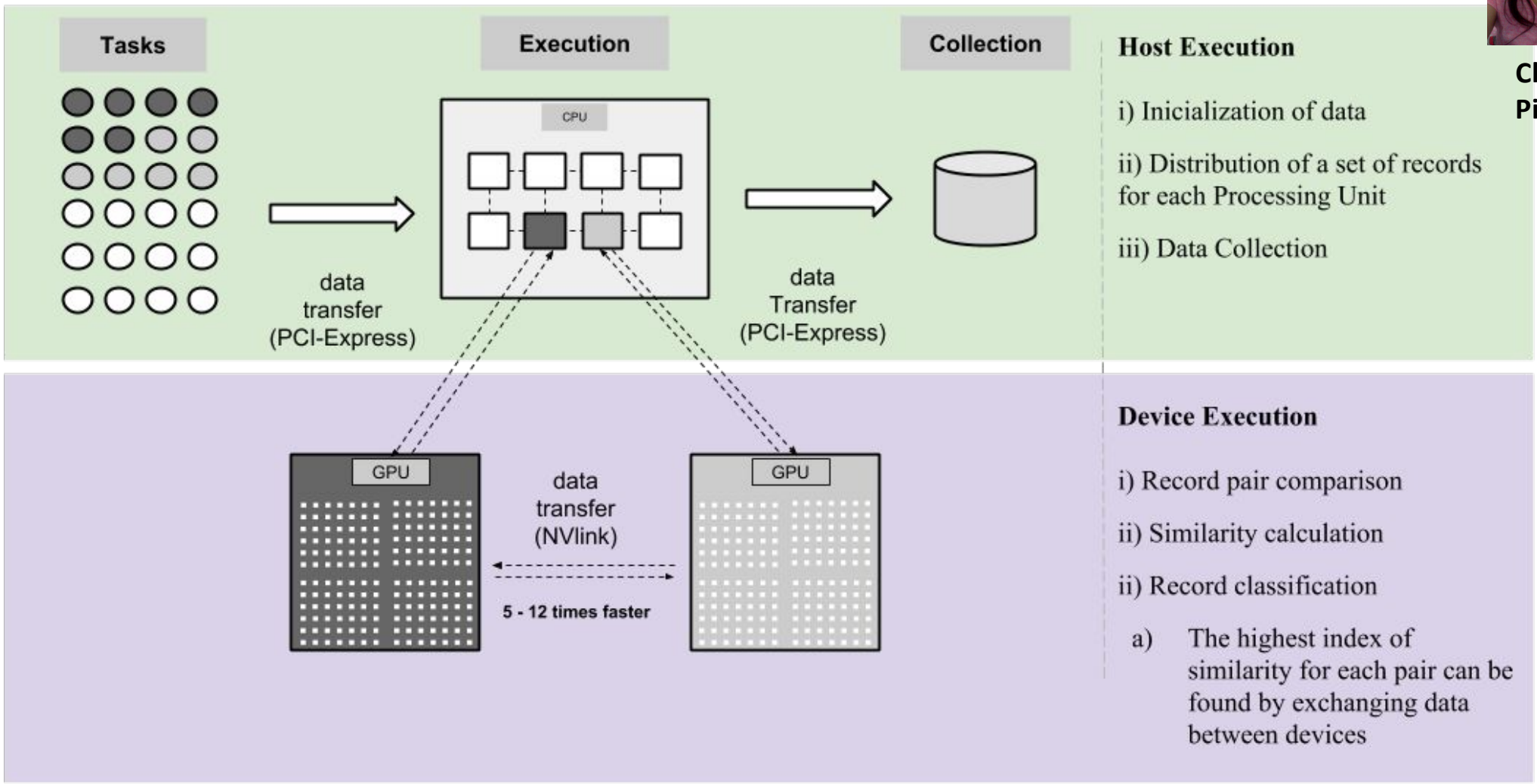
region 1	region 2	region 3	region 4
serial	50% parallel	comms	90% parallel



An example of a heterogeneous multi-core and multi-GPU system.

✓ Exploring hybrid parallel systems for probabilistic record linkage

Clícia Pinto



Host Execution

i) Inicialization of data

ii) Distribution of a set of records for each Processing Unit

iii) Data Collection

Device Execution

i) Record pair comparison

ii) Similarity calculation

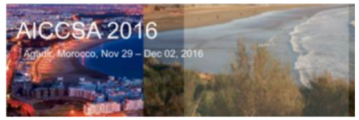
ii) Record classification

a) The highest index of similarity for each pair can be found by exchanging data between devices

Auto-Tuning TRSM with an Asynchronous Task Assignment Model on Multicore, Multi-GPU and Coprocessor systems

Clicia Pinto and Marcos Barreto
Laboratório de Sistemas Distribuídos
Universidade Federal da Bahia (UFBA)
Salvador – Bahia – Brazil
{cliciasp,marcosb}@ufba.br

Murilo Boratto
Núcleo de Arquitetura de Computadores
e Sistemas Operacionais
Universidade do Estado da Bahia (UNEB)
Salvador – Bahia – Brazil
muriloboratto@uneb.br



SpringerLink

The Journal of Supercomputing
pp 1–13 | Cite as

Exploring hybrid parallel systems for probabilistic record linkage

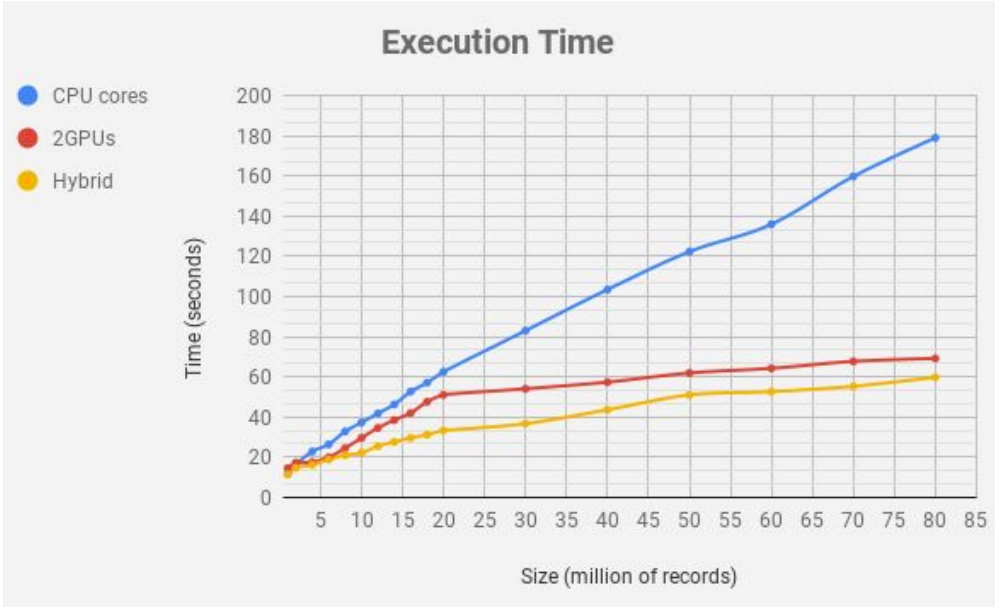
Authors Authors and affiliations

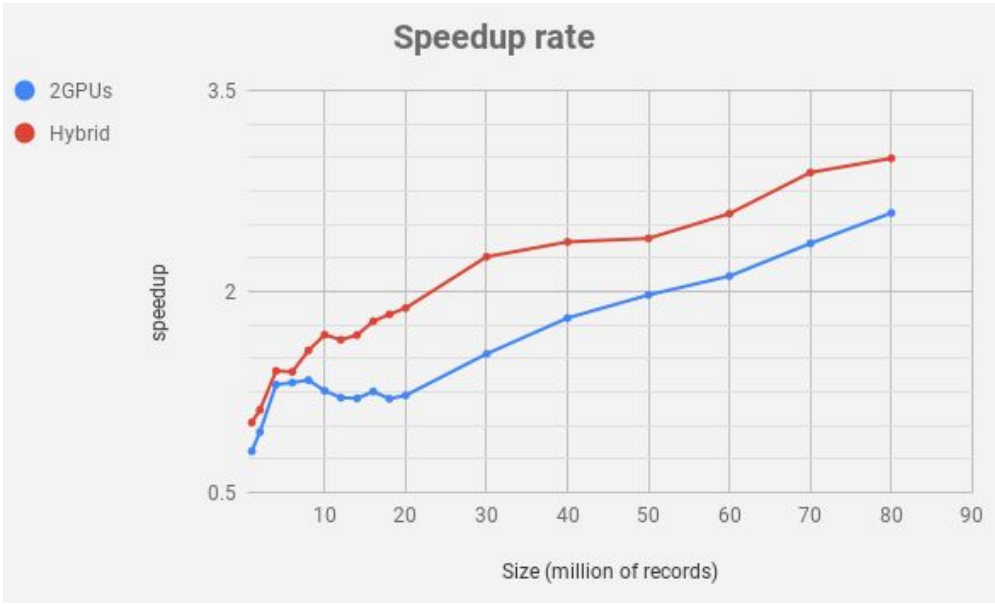
Murilo Boratto, Pedro Alonso, Clicia Pinto, Pedro Melo, Marcos Barreto, Spiros Denaxas

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



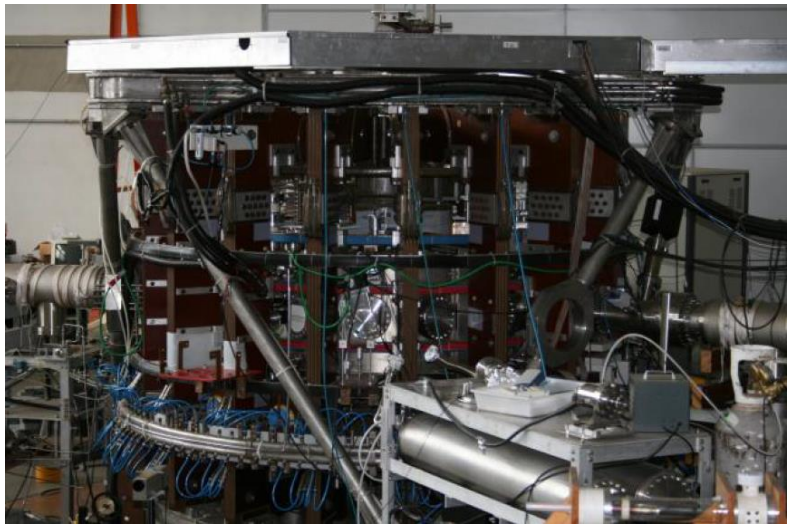


hybrid parallel computing

Elizabeth Gomes

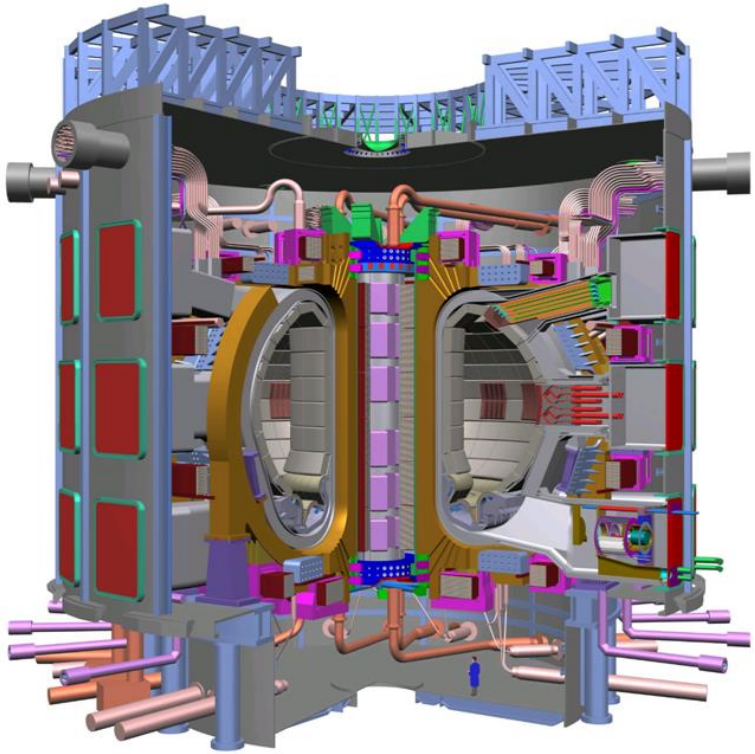
✓ **High-performance plasma Physics**

Tokamak no Brasil
(câmara toroidal com bobinas magnéticas)



TCABR, departamento de Física Aplicada da USP

Fonte: (TCABR.. . , 2017)

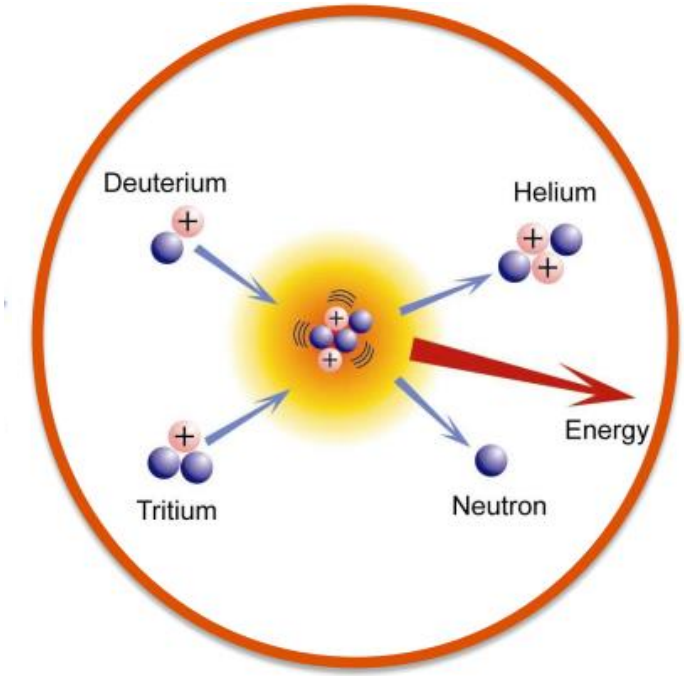


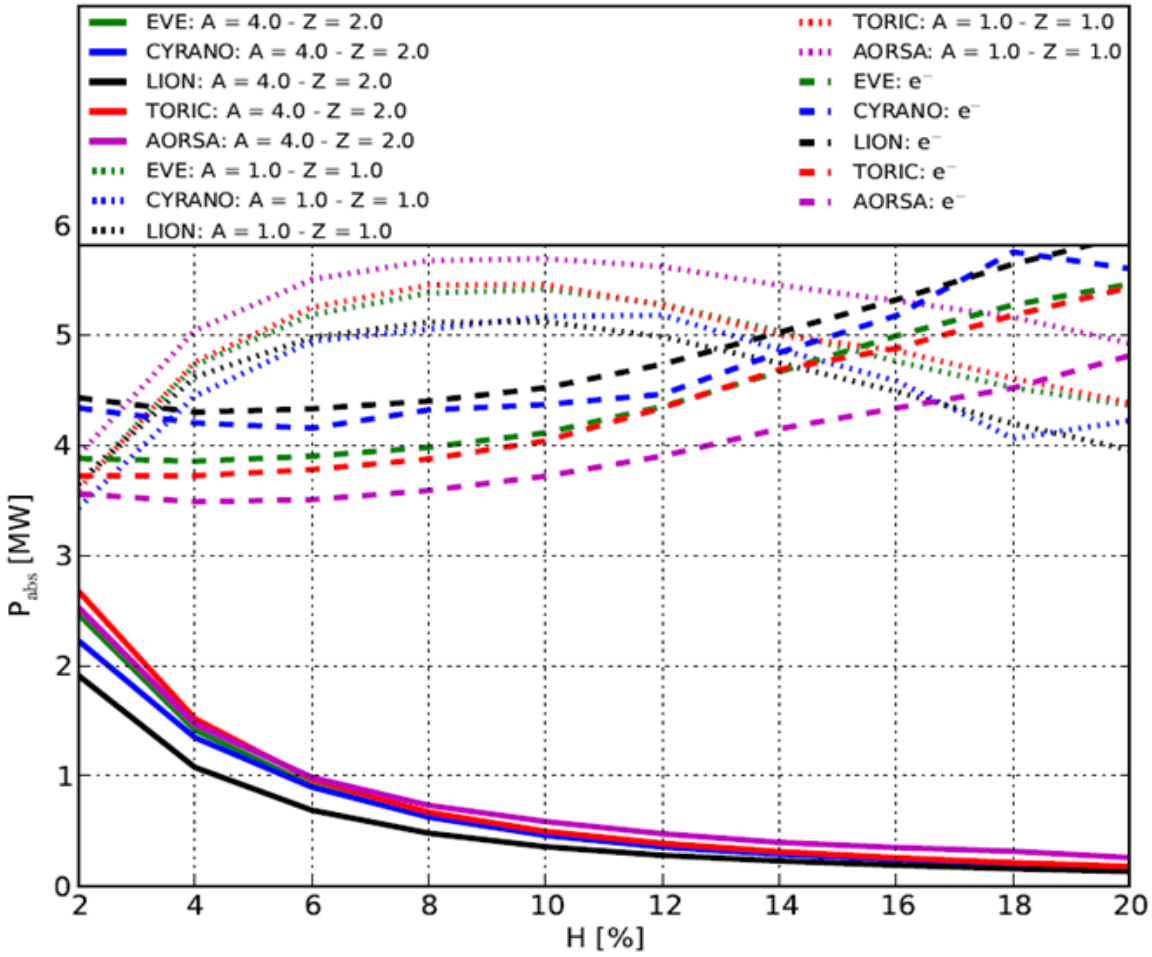
Fonte: (BUTLER, 2013).

✓ **High-performance plasma Physics**

✓ CYRANO – Cyclotron Resonance Analysis with No Obfuscation.

✓ Plasma heating by RF waves.
✓ Isotopes of hydrogen.





REVISTA Mundi
Engenharia, Tecnologia e Gestão
INSTITUTO FEDERAL Paraná

Figura 3 – Relação do consumo do tempo das funções GENERA3 e OUTPOW em relação ao tempo total, com 280 e 480 elementos. Código original.

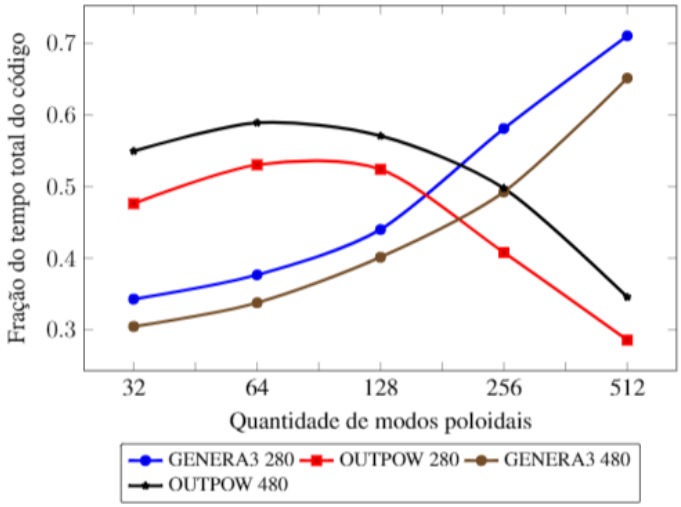
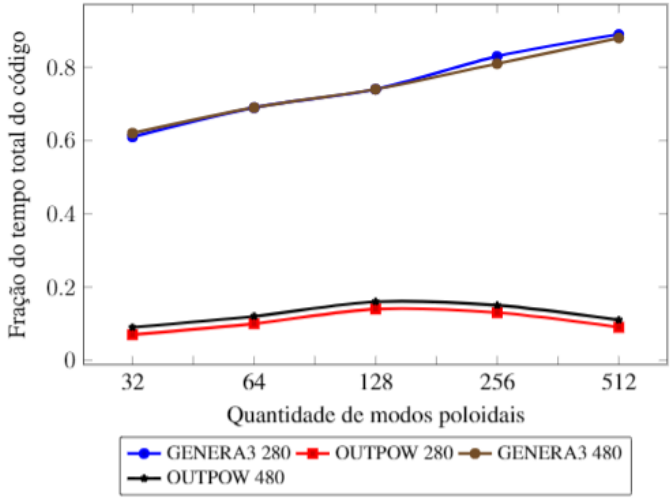


Figura 4 – Gráfico do consumo do tempo pelas funções GENERA3 e OUTPOW em relação ao tempo total, com 280 e 480 elementos. Código com OUTPOW OpenMP e OpenBLAS.



APRIMORAMENTO DO CÓDIGO CYRANO DE SIMULAÇÃO DE AQUECIMENTO DE PLASMA POR ONDAS RF UTILIZANDO TÉCNICAS DE PROCESSAMENTO PARALELO

IMPROVEMENT OF THE CYRANO CODE FOR PLASMA HEATING SIMULATION BY RF WAVES USING PARALLEL PROCESSING TECHNIQUES

Anusio Menezes Correia [1]
Ernesto Augusto Lerche [2]
Dirk Van Eester [3]
Gesil Sampaio Amarante Segundo [4]
Esbel Tomas Valero Orellana [5]

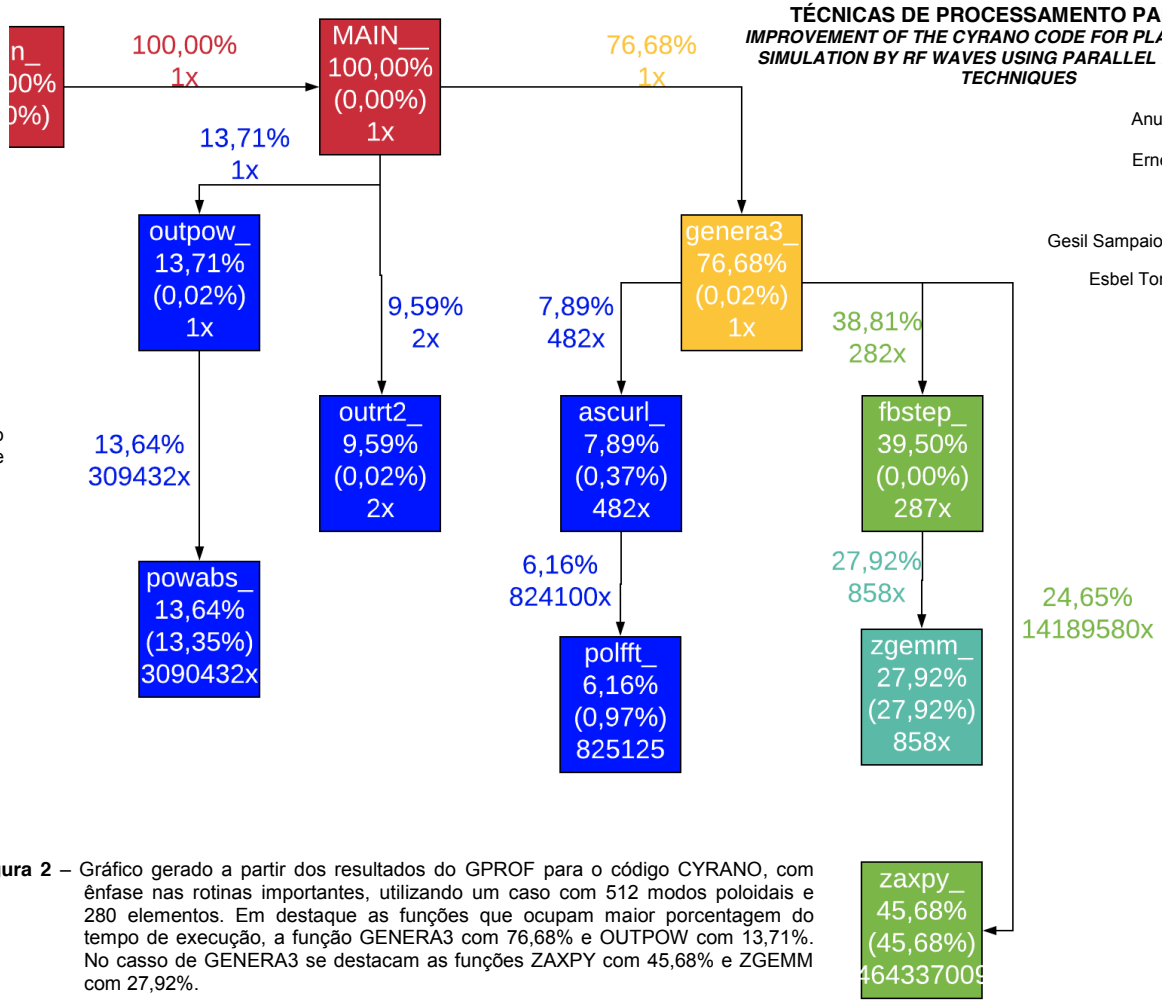


Figura 2 – Gráfico gerado a partir dos resultados do GPROF para o código CYRANO, com ênfase nas rotinas importantes, utilizando um caso com 512 modos poloidais e 280 elementos. Em destaque as funções que ocupam maior porcentagem do tempo de execução, a função GENERA3 com 76,68% e OUTPOW com 13,71%. No casso de GENERA3 se destacam as funções ZAXPY com 45,68% e ZGEMM com 27,92%.

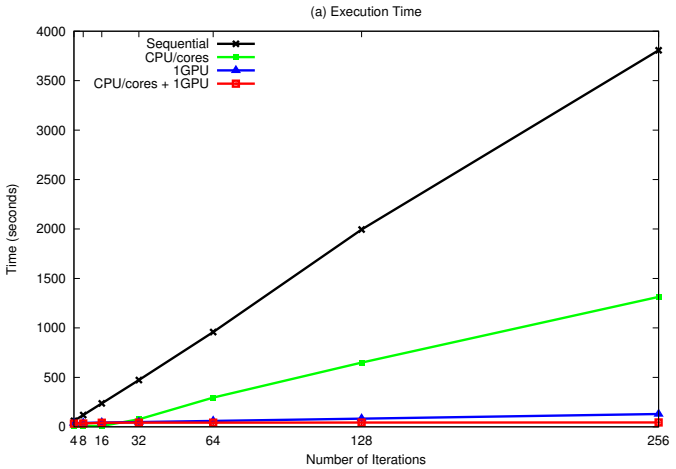
hybrid parallel computing

International Conference on Computational Science and Its Applications
ICCSA 2017: Computational Science and Its Applications – ICCSA 2017 pp 439-451 | Cite as

Accelerating Docking Simulation Using Multicore and GPU Systems

Authors | Authors and affiliations

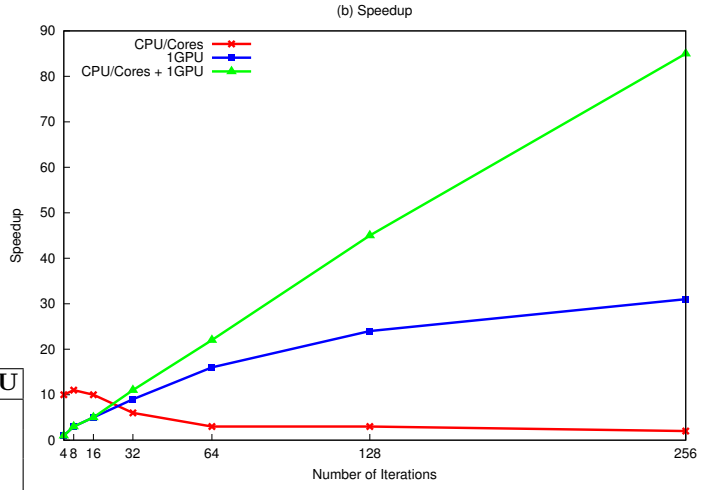
Everton Mendonça, Marcos Barreto, Vinícius Guimarães, Nelci Santos, Samuel Pita, Murilo Boratto



Everton Mendonça

Table 1. Costliest Functions in Autodock

% Time Spent	Execution Time (seconds)	Function/Routine
38.61	57.51	$eintcal()$
31.47	46.88	$trilinterp()$
4.12	6.13	$torsion()$
3.32	4.95	$snorm()$
2.96	4.41	$RealVector :: clone()const$



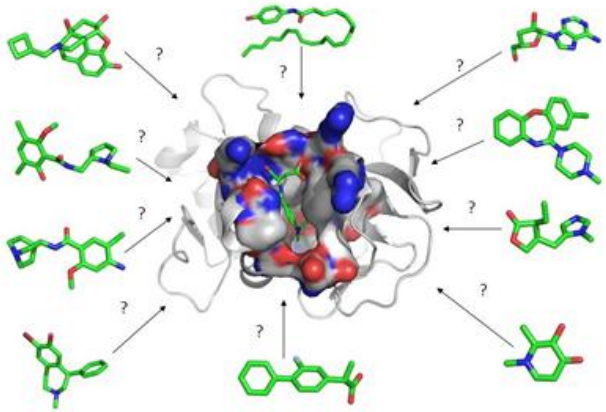


Table 2. Execution time (s) obtained with different parameter values.

Iterações	Sequencial	CPU/Cores	1GPU	CPU/Cores + 1GPU
1	15.97	**13.50**	37.26	42.12
2	30.33	**13.20**	37.72	35.26
4	60.25	**10.17**	38.40	33.79
8	119.66	**09.98**	39.70	34.27
16	237.61	**09.48**	43.01	41.94
32	473.04	76.79	**48.16**	**42.30**
64	958.89	295.24	**59.91**	**42.82**
128	1995.260	649.78	**82.85**	**43.42**
256	3807.390	1314.63	**129.04**	**44.56**

hybrid parallel computing

Support for bioinformatics applications through volunteer and scalable computing frameworks

Felipe Gutierrez, Danilo Azevedo and Marcos Barreto
Distributed Systems Lab. (LaSiD)
Computer Science Department
Federal University of Bahia (UFBA)
Salvador, Brazil 40110–170
Email: felipe.o.gutierrez@gmail.com,
(azevedospider,marcoseb)@dcc.ufba.br

Rodrigo Zucoloto
Population Genetics and Molecular Evolution Lab. (GENEV)
Institute of Biology
Federal University of Bahia (UFBA)
Salvador, Brazil 40110–170
Email: rbz@ufba.br

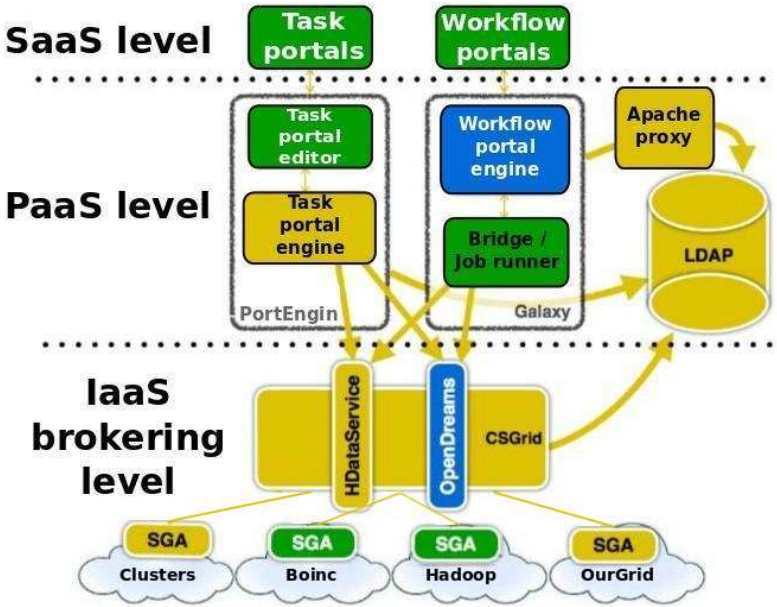


Fig. 1. mc^2 platform architecture.

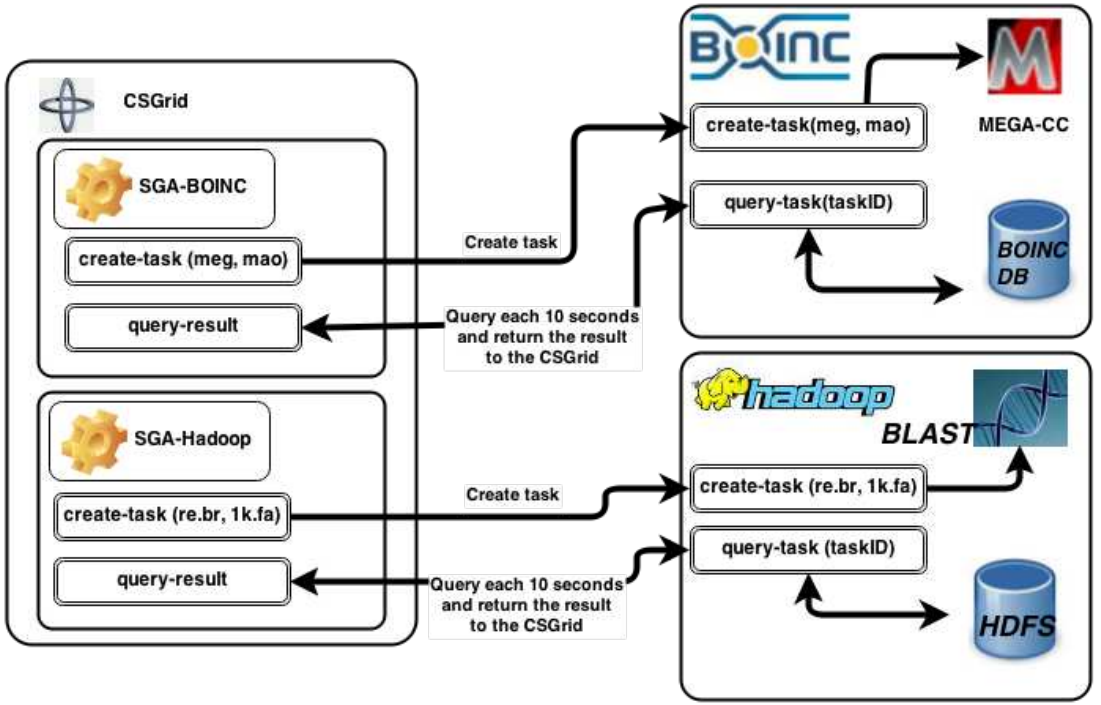
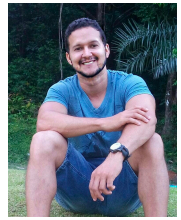


Fig. 3. Example of a workflow in mc^2.

Felipe Gutierrez



Danilo Azevedo

TABLE II. MODEL SELECTION ALGORITHM IN MEGA-CC (HETEROGENEOUS CASE).

Tasks	3	9	15	18	21
Mega seq.	1600.50	4801.50	8002.50	9603.00	11203.50
mc^2 + Mega	624.00	1835.00	3048.00	3640.00	4232.00
Accel.	2.57	2.62	2.63	2.64	2.65

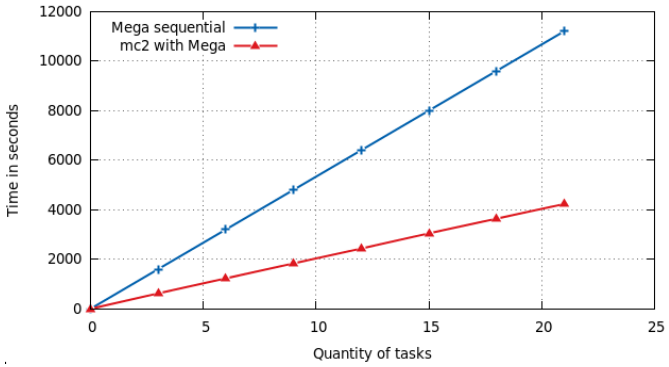


Fig. 5. Model selection algorithm in MEGA-CC (heterogeneous case).

TABLE III. RESULTS FOR CLOUDBURST APPLICATION.

Application	1 PC	PC + 1 iMAC	PC + 2 iMAC	PC + 3 iMAC
CloudBurst	101.685	56.254	47.724	42.193
mc^2 + CloudBurst	113.5	78.1	58.3	53.5

TABLE IV. RESULTS FOR BLAST IN HADOOP.

Nodes	1	2	3	4
BLAST	911.34	460.43	289.59	235.79
BLAST on mc^2	924.6	469.3	293.4	240.96

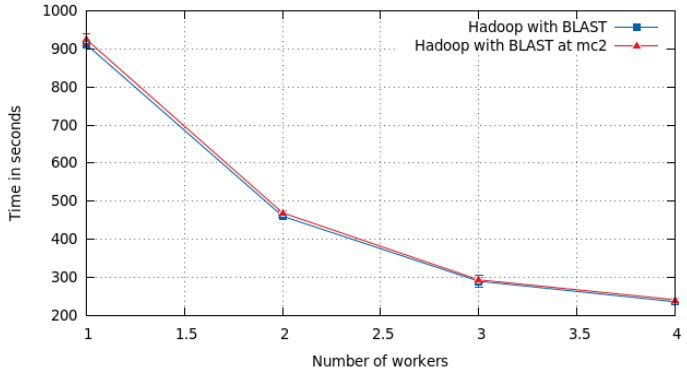


Fig. 7. Comparison of BLAST algorithm in Hadoop and mc^2.

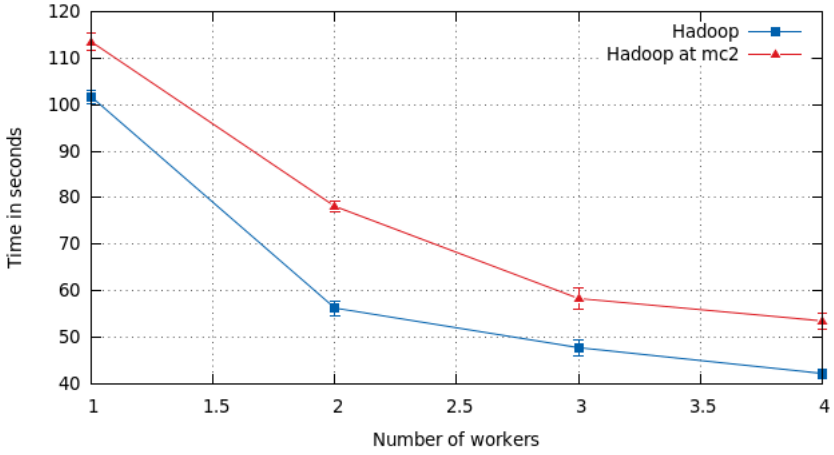
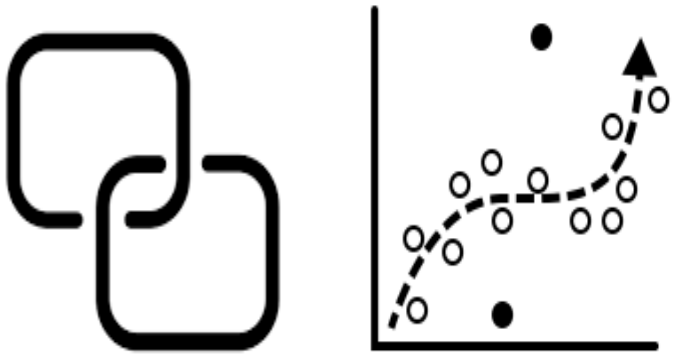


Fig. 6. CloudBurst application in Hadoop and in mc^2.

What we do

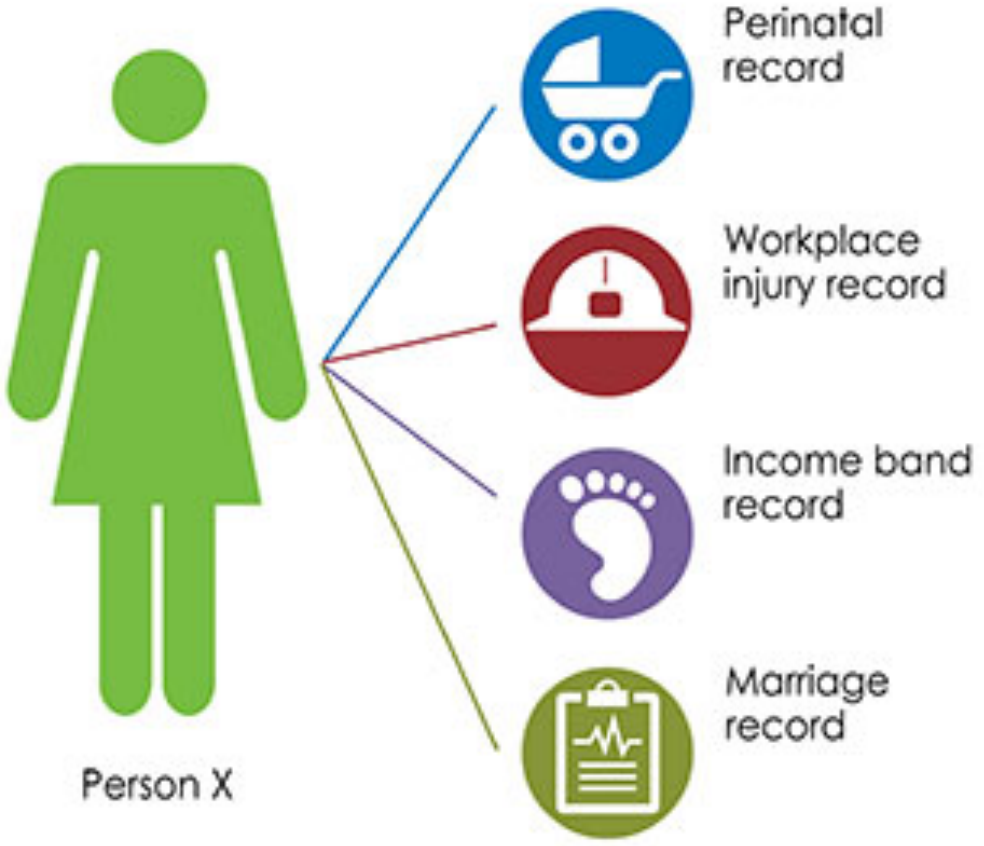


big data
linkage &
analytics

data linkage & analytics

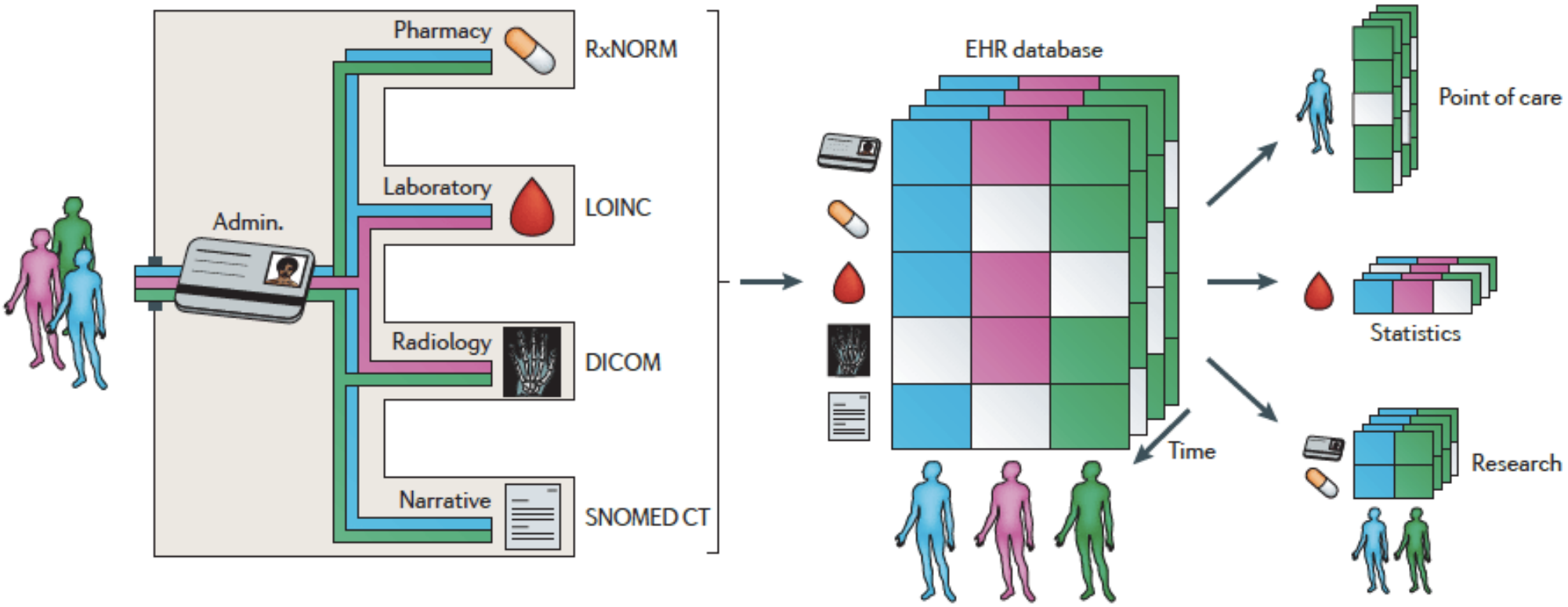
✓ **Why do we need data linkage?**

✓ To aggregate data from multiple sources that presumably pertain to the same real world entity.

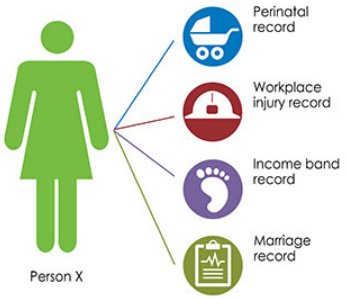


data linkage & analytics

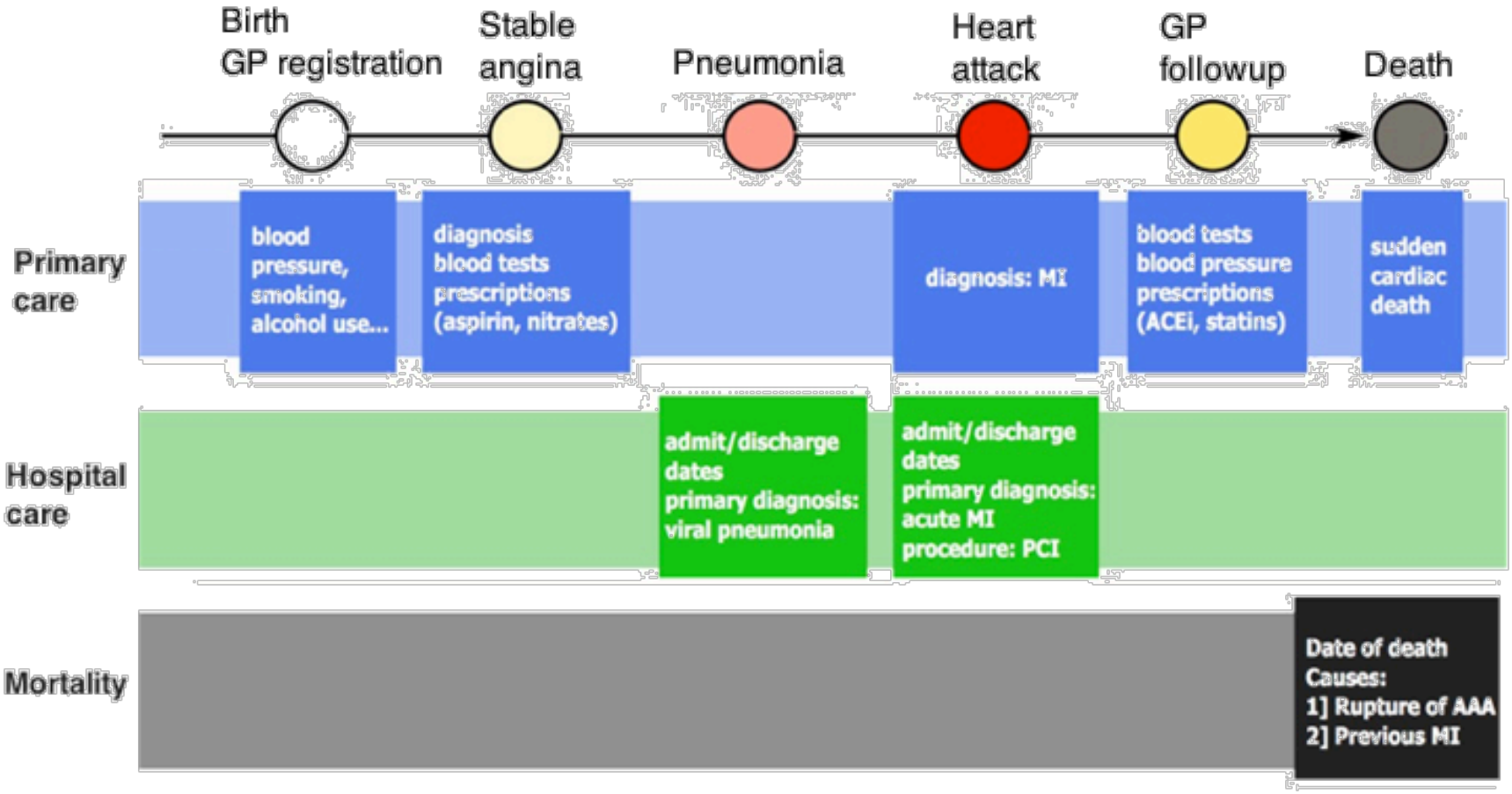
✓ **Example scenario: electronic health records (EHR)**



data linkage & analytics



✓ **EHR: patient pathway**



CALIBER

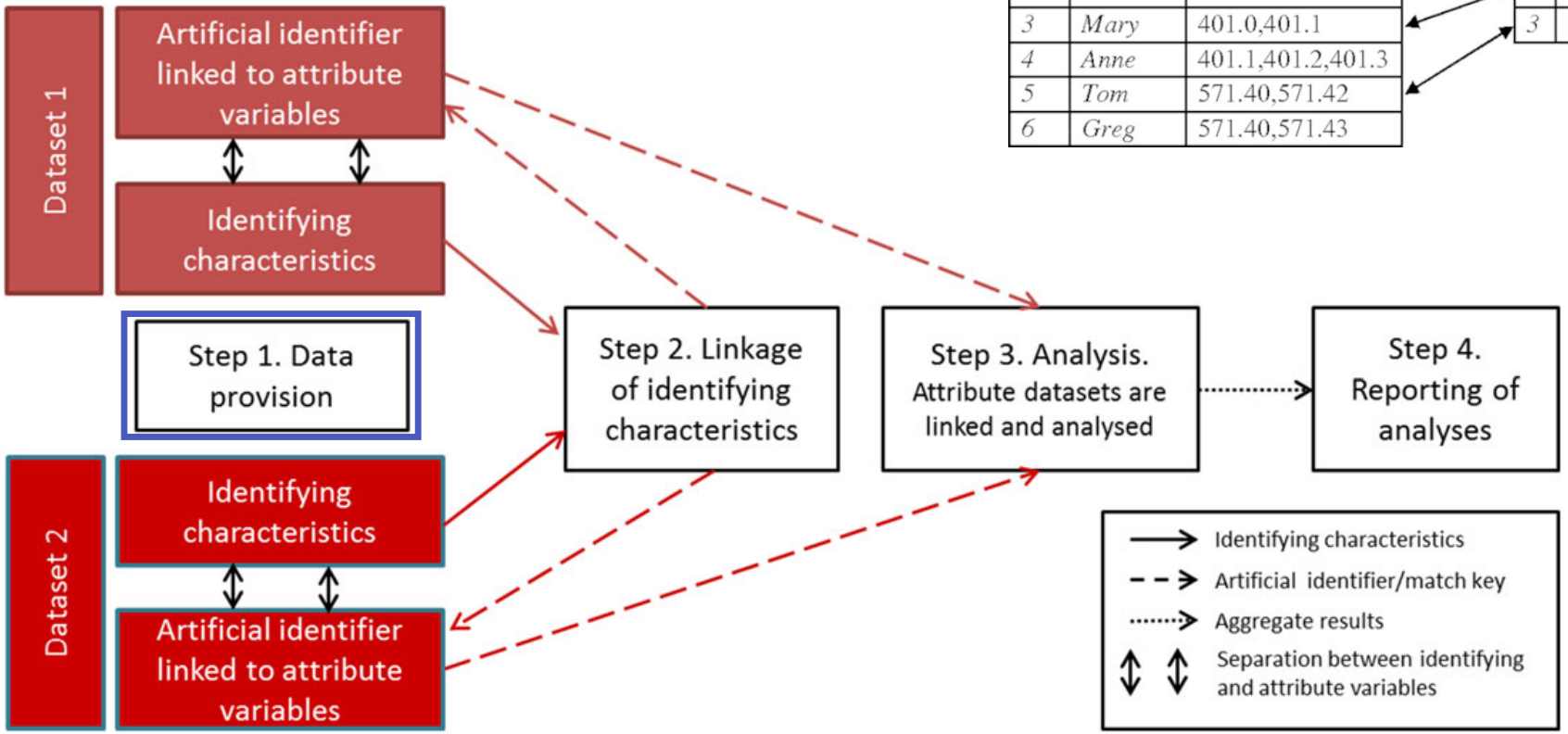
UCL

data linkage & analytics

✓ Data linkage pipeline: basic pathway

Identified EMR data (P)		
i	ID	ICD9
1	Jim	493.00
2	Jack	493.00
3	Mary	401.0,401.1
4	Anne	401.1,401.2,401.3
5	Tom	571.40,571.42
6	Greg	571.40,571.43

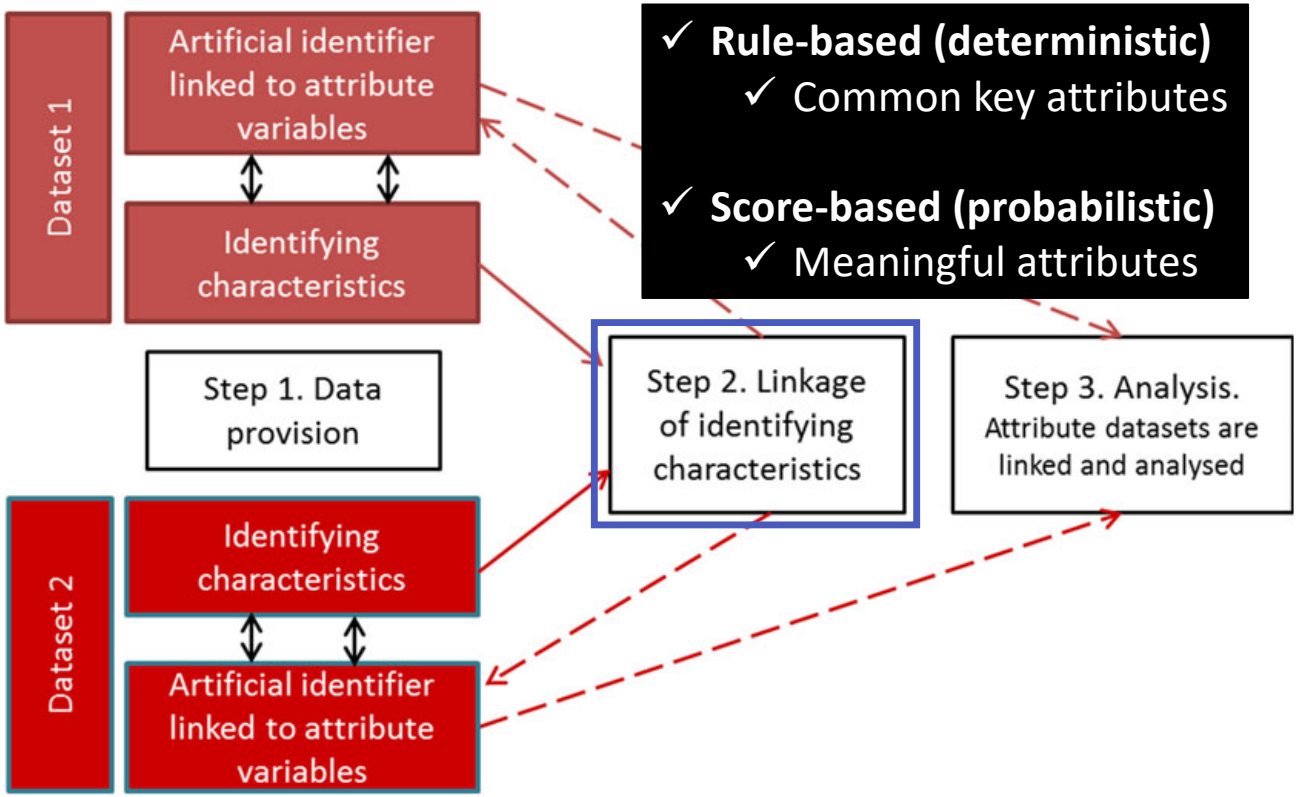
De-identified Research data (S)		
j	ICD9	DNA
1	493.00	CT…A
2	401.0,401.1	AC…T
3	571.40,571.42	GC…A



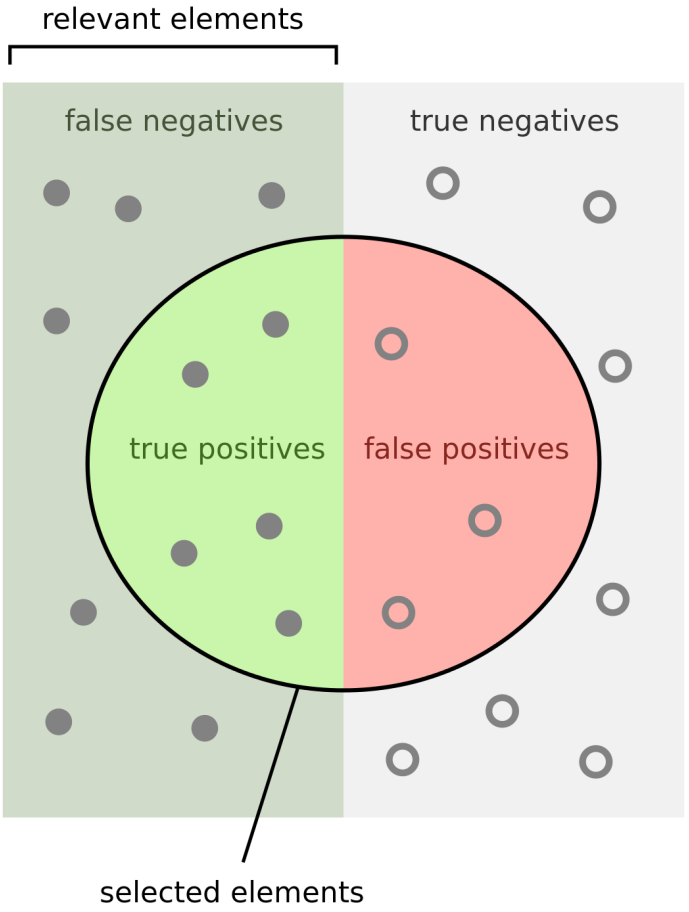
GUILD: GUidance for Information about Linking Data sets
Journal of Public Health , doi:10.1093/pubmed/fdx037

data linkage & analytics

✓ Data linkage pipeline: pairwise comparison

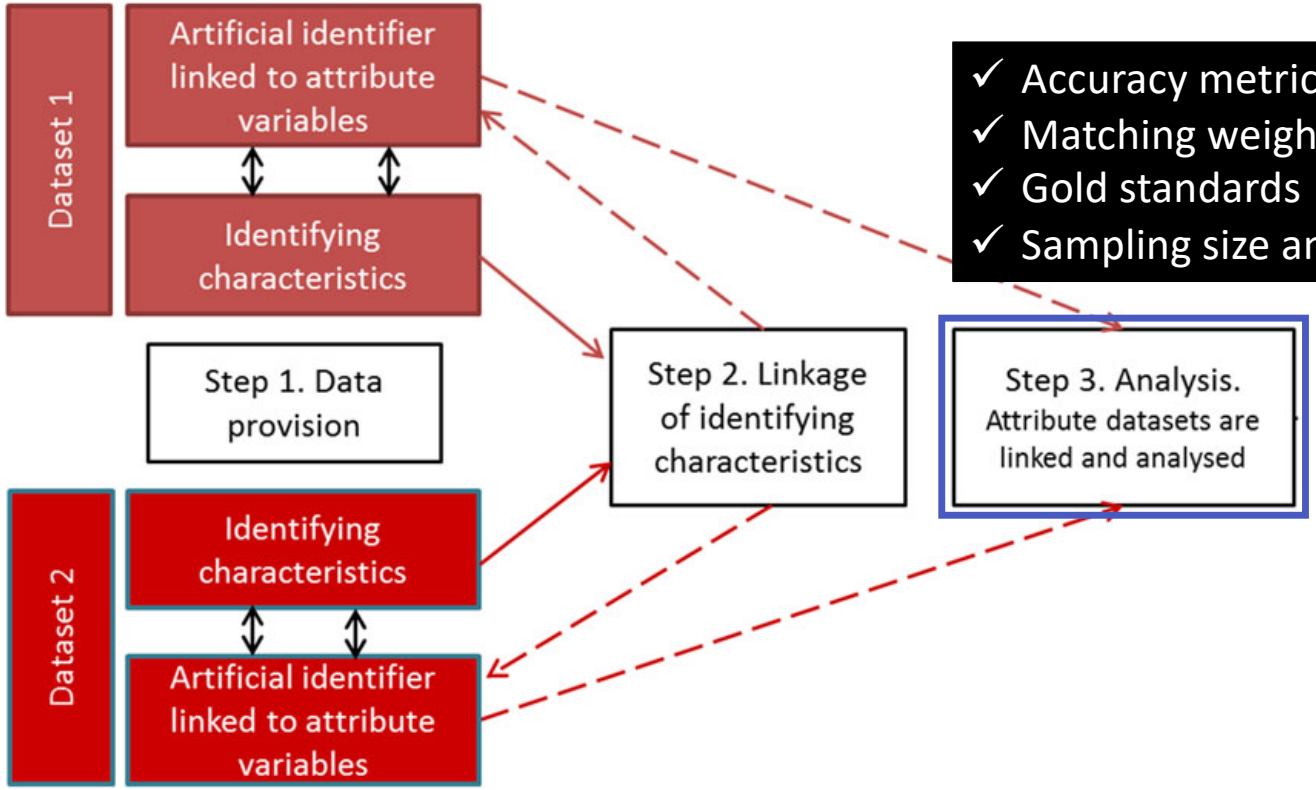


GUILD: GUidance for Information about Linking Data sets
Journal of Public Health , doi:10.1093/pubmed/fdx037

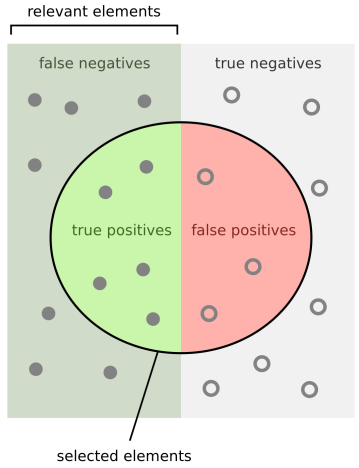


data linkage & analytics

✓ Data linkage pipeline: accuracy assessment



- ✓ Accuracy metrics
- ✓ Matching weights
- ✓ Gold standards
- ✓ Sampling size and expressiveness



How many selected items are relevant?

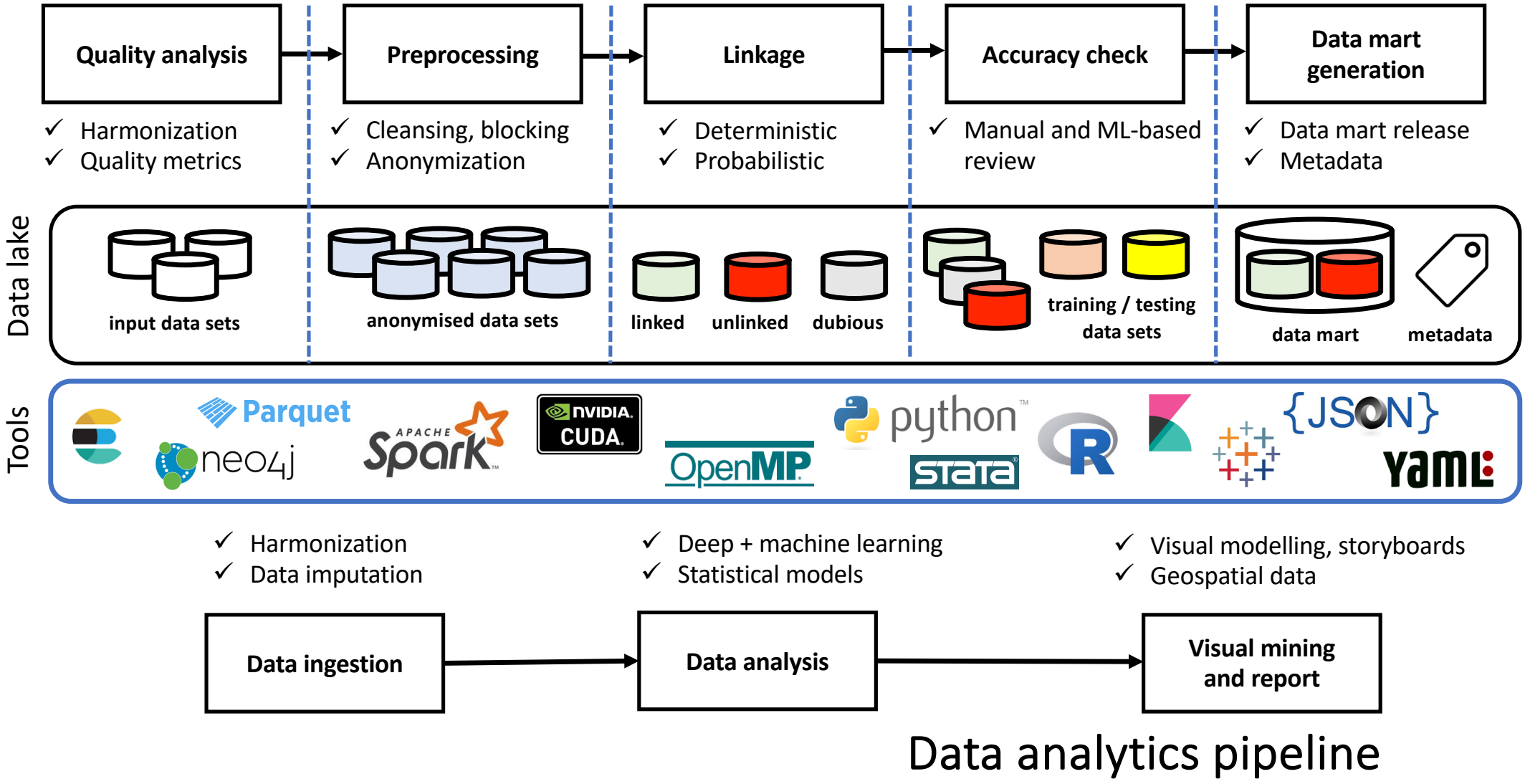
$$\text{Precision} = \frac{\text{true positives}}{\text{selected elements}}$$

How many relevant items are selected?

$$\text{Recall} = \frac{\text{true positives}}{\text{relevant elements}}$$

GUILD: GUidance for Information about Linking Data sets
Journal of Public Health , doi:10.1093/pubmed/fdx037

AtyImo – Data linkage platform

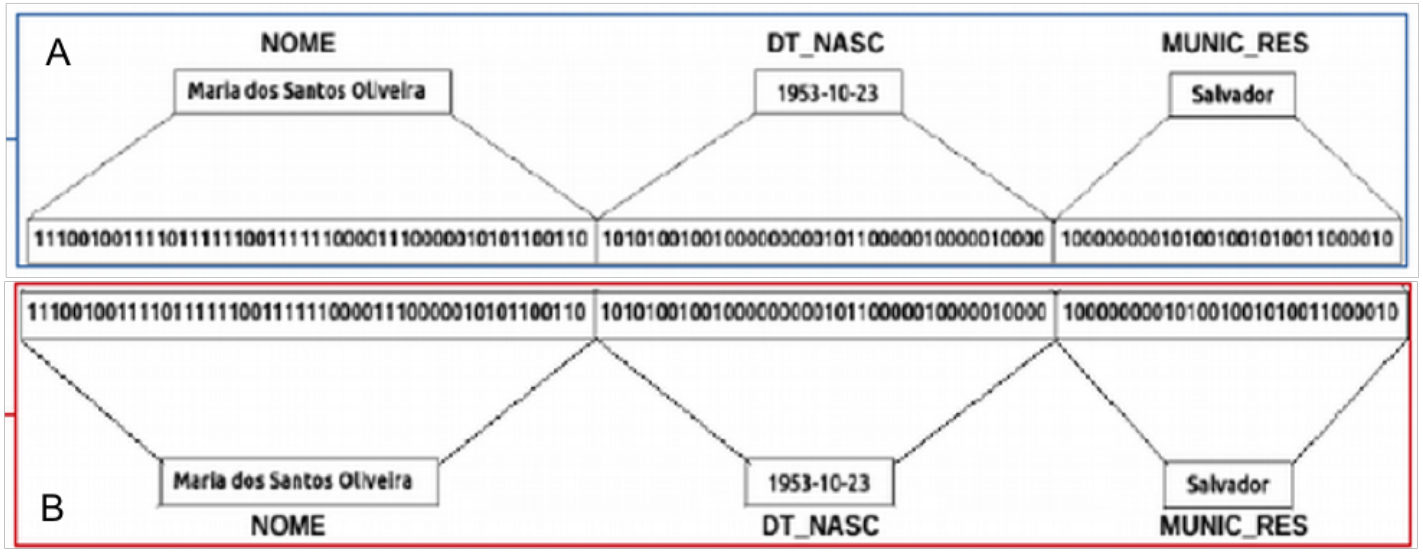


Data analytics pipeline

AtyImo

✓ AtyImo: probabilistic linkage

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



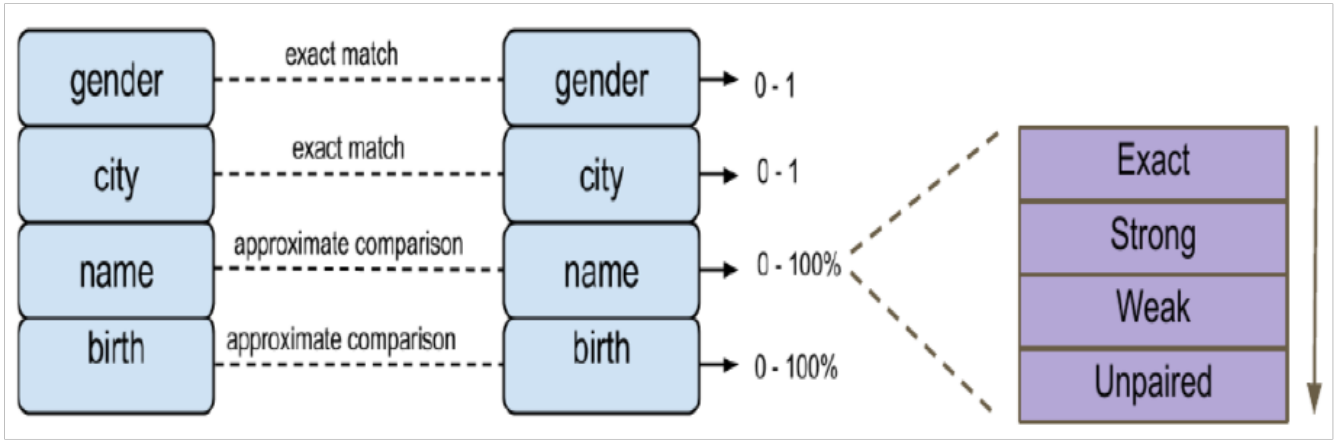
$$D_{a,b} = \frac{2h}{|a| + |b|} = [0, 1]$$

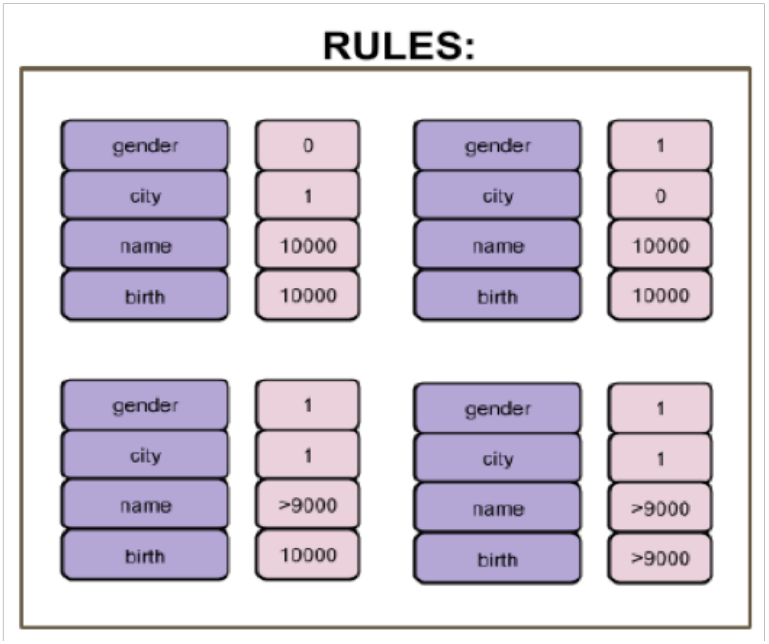
h = number of 1's at same position in both Bloom filters
a = number of 1's in Bloom filter A
b = number of 1's in Bloom filter B

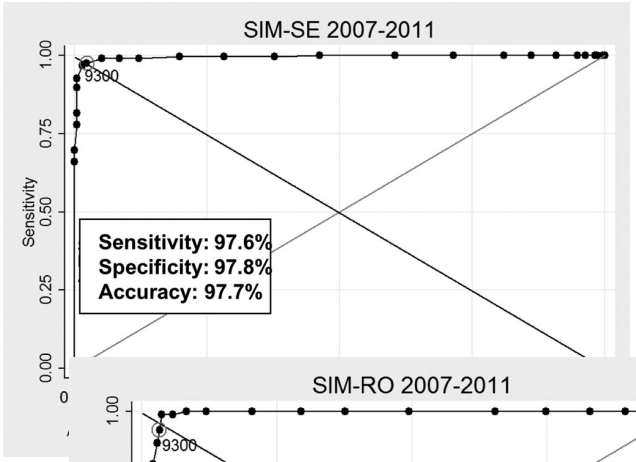
AtyImo

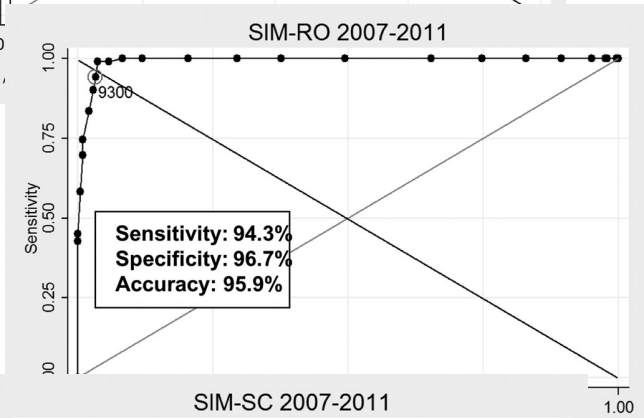
✓ **AtyImo: hybrid (probabilistic + deterministic) linkage**

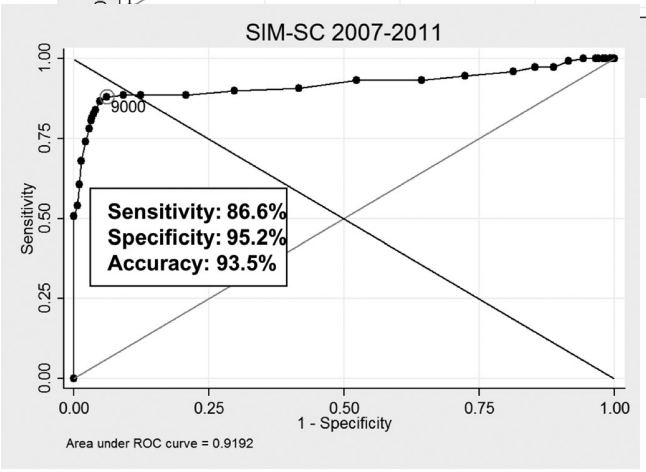
- **Hybrid approach**: individual comparison of attributes based on different rules











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

AtyImo

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]

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

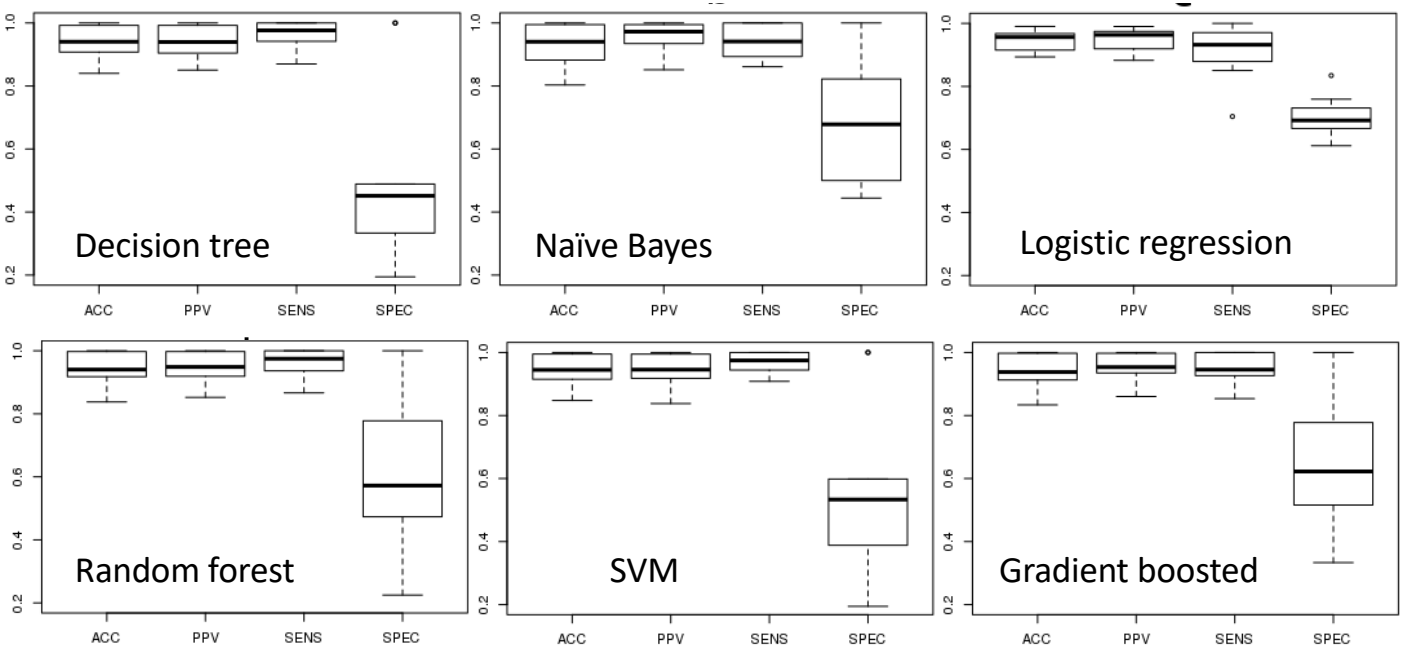
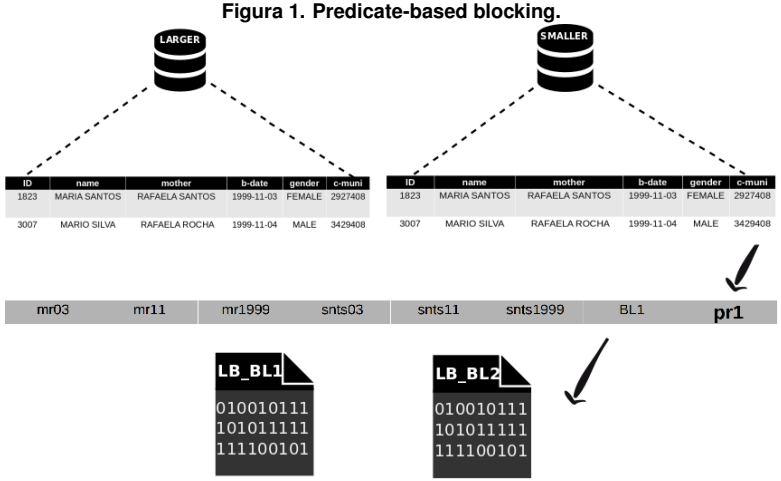
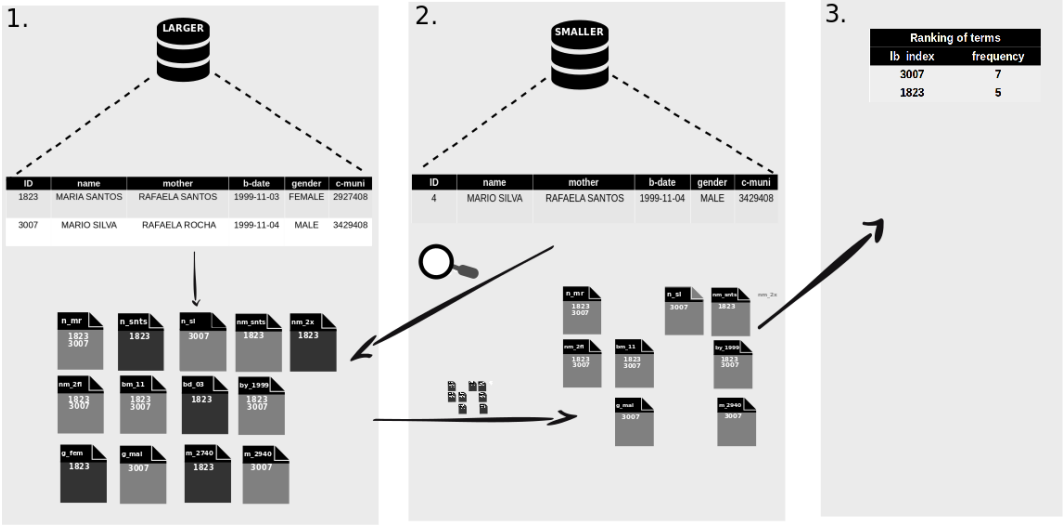


Figura 1. Predicate-based blocking.



$(birthday \vee birthmonth \vee birthyear)) \vee (lastname \wedge (birthday \vee birthmonth \vee birthyear))$ and the second as $pr2 = ((firstname \wedge firstmothersname) \wedge (birthday \vee birthmonth \vee birthyear)) \vee ((lastname \wedge lastmothersname) \wedge (birthday \vee birthmonth \vee birthyear))$.

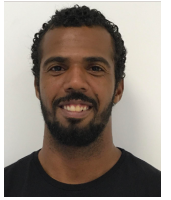
Figura 2. Term frequency-based approach used in Atylmo.



Latin America Data Science Workshop
AUGUST 27TH – VLDB 2018 WORKSHOP – RIO DE JANEIRO, BRAZIL

Applying term frequency-based indexing to improve scalability and accuracy of probabilistic data linkage

Robespierre Pita[1,2], **Luan Menezes**[1,2], **Marcos E. Barreto**[1,2]



Robespierre Pita

Tabela 1. Gold standard data set used for validation.

SIM	6,458 records
SINASC	13,046 records
Total of comparisons	84,251,068 records
Expected true positives	3,030 records

Tabela 2. Size of generated blocks for each indexing technique

method	**predicate 1**		**predicate 2**		**term frequecy**	
database	***sb***	***lb***	***sb***	***lb***	***sb***	***lb***
min	1	1	1	1	1	100
med	24	51	2	2	1	100
mean	43	88.38	8.289	11.57	1	100
max	1855	41528	87	611	1	100

Tabela 3. Results of each indexing techinique used.

	predicate 1	**predicate 2**	**term frequency**
true matches retrieved	2,382	3,018	3,020
number of blocks	5,806	6,432	6,458
number of comparisons	44,406,049	29,111,755	645,800
reduction ratio	0.472	0,654	0,992
pair completness	0.786	0.996	0.996

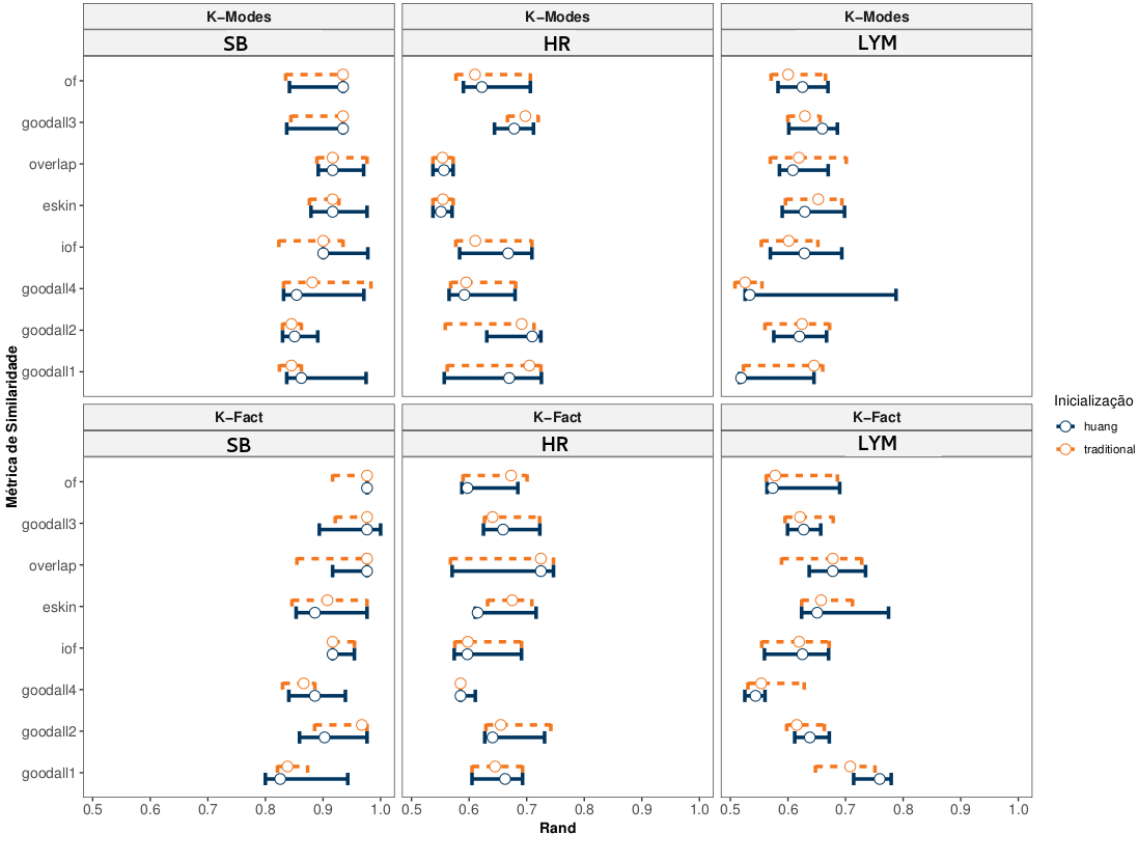
K-fact: Using the Frequency Factor for Clustering Categorical Data

TABLE III
SIMILARITY MEASURES FOR CATEGORICAL DATA.

Measure	Per-attribute similarity $S_j(X_j, Y_j)$	
Overlap	$= \begin{cases} 1 \\ 0 \end{cases}$	If $X_j = Y_j$ otherwise
Eskin	$= \begin{cases} 1 \\ \frac{v_j^2}{v_j^2 + 2} \end{cases}$	If $X_j = Y_j$ otherwise
Goodall1	$= \begin{cases} 1 - \sum_{v \in V} p_j^2(v) \\ 0 \end{cases}$	If $X_j = Y_j$ otherwise
Goodall2	$= \begin{cases} 1 - \sum_{v \in V} p_j^2(v) \\ 0 \end{cases}$	If $X_j = Y_j$ otherwise
Goodall3	$= \begin{cases} 1 - p_j^2(X_j) \\ 0 \end{cases}$	If $X_j = Y_j$ otherwise
Goodall4	$= \begin{cases} p_j^2(X_j) \\ 0 \end{cases}$	If $X_j = Y_j$ otherwise
OF	$= \begin{cases} 1 \\ \frac{1}{1 + \log \frac{N}{f_j(X_j)} \times \log \frac{N}{f_j(Y_j)}} \end{cases}$	If $X_j = Y_j$ otherwise
IOF	$= \begin{cases} 1 \\ \frac{1}{1 + \log f_j(X_j) \times \log f_j(Y_j)} \end{cases}$	If $X_j = Y_j$ otherwise

TABLE IV
DESCRIPTIVE ANALYSIS OF THE DATASETS USED TO EVALUATE OUR MODEL IN COMPARISON TO K-MODES.

		UCI datasets			Simulated datasets			
		SB	**HR**	**LYM**	**n100**	**n250**	**n500**	**n1K**
N	before	47	132	148	100	250	500	1K
	after	47	132	148	100	250	500	1K
p	before	35	6	19	5	5	5	5
	after	21	4	19	5	5	5	5
v_j	before	2.11	25	3.31	3.2	3.2	3.2	3.2
	after	2.76	3.6	3.31	3.2	3.2	3.2	3.2
na?	before	no	no	no	no	no	no	no
	after	no	no	no	no	no	no	no
k	before	4	3	4	3	3	3	3
	after	4	3	4	3	3	3	3



Projects and collaborations

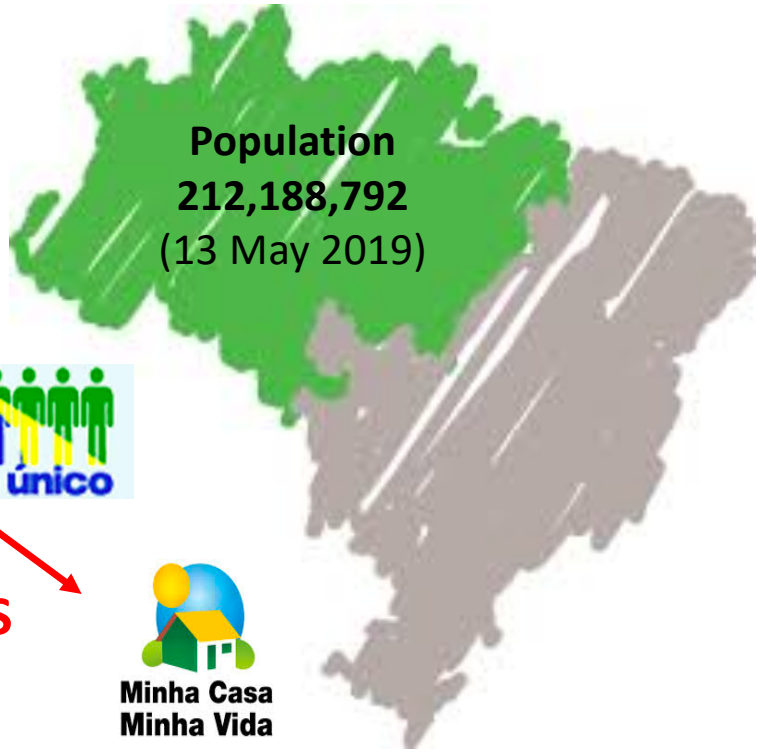
PARTNERS

SPONSORS

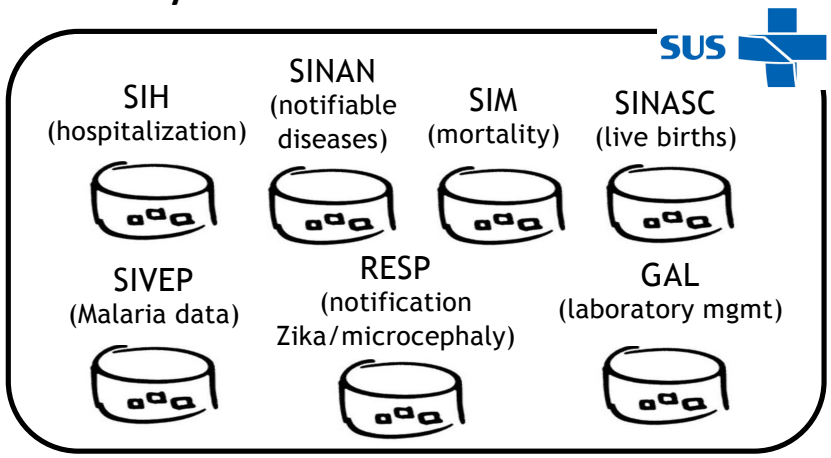
100 Million Cohort

Social programmes

+10 other programmes



Public health system



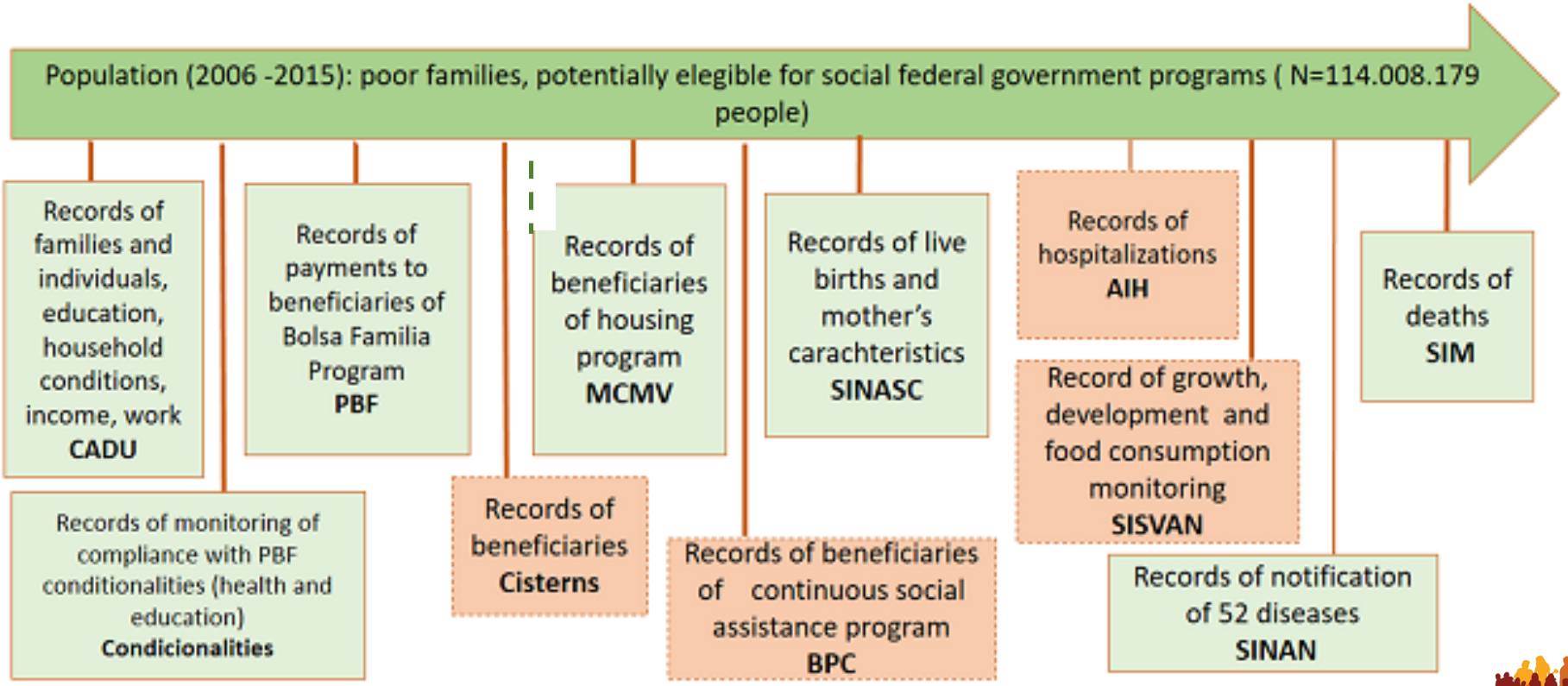
- ✓ Individuals registered in CADU
- ✓ Payments from Bolsa Família (cash transfers) + MCMV (housing)
- ✓ Period: 2006 – 2015
- ✓ **114 million individuals**
- ✓ ≦ 5,000 variables / individual

Deterministic linkage Probabilistic linkage

cidacs
Centro de Integração de Dados e Conhecimentos para Saúde
FIOCRUZ

The 100 Million Brazilians Cohort

CADASTRO ÚNICO – CADU
(Unified Registry for Social programs - From 2004)



Coorte de 100 milhões de brasileiros

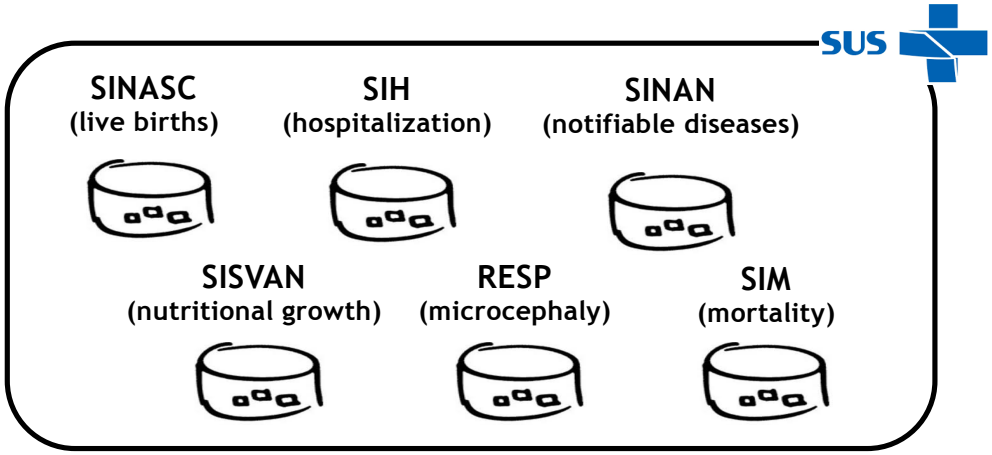
PLATAFORMA ZIKA

Zika Platform

This platform aims to improve scientific knowledge about the disease and support the adoption of more appropriate public health measures to deal with the triple epidemic caused by Zika, Dengue and Chikungunya.

cidacs
Centro de Integração de Dados e Conhecimentos para Saúde
FIOCRUZ

- ✓Birth cohort, 2001 – 2030
- ✓≅80 million records.



Epidemiology | **Research** | **Collaboration** | **Data curation** | **Open Science**

Integrating socioeconomic and healthcare data to combat malaria

✓i) data integration and ii) epidemics forecasting.

SIVEP
- ✓ Coverage: 2003-2018
- ✓ Records: 5,340,564
- ✓ Attributes: 52

SIM
- ✓ Coverage: 2003-2018
- ✓ Records: 1,004
- ✓ Attributes: 37

SINAN
- ✓ Coverage: 2003-2018
- ✓ Records: 46,170
- ✓ Attributes: 20

Climate
- ✓ Coverage: 2003-2018
- ✓ Records: 5,570
- ✓ Attributes: 5

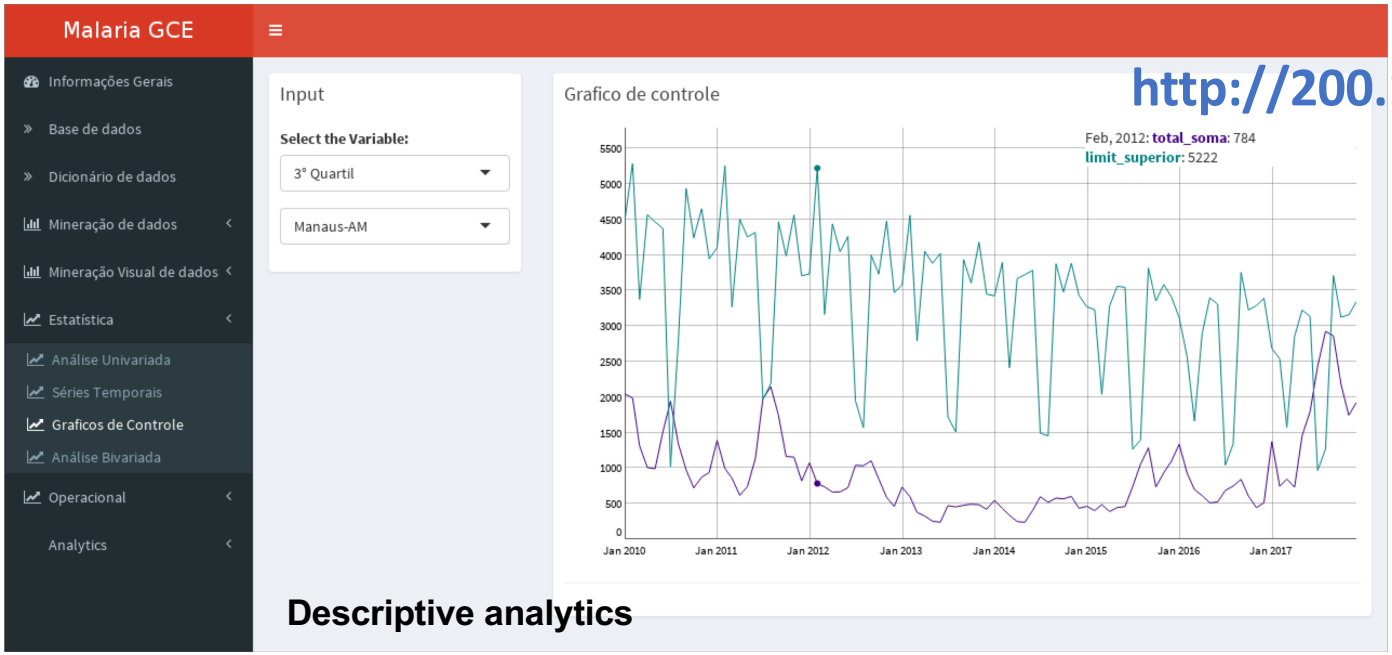
Vector control (2016-2018)

- ✓ Laboratory
- ✓ Records: 75
- ✓ Attributes: 11

- ✓ Hotspots
- ✓ Records: 325
- ✓ Attributes: 26

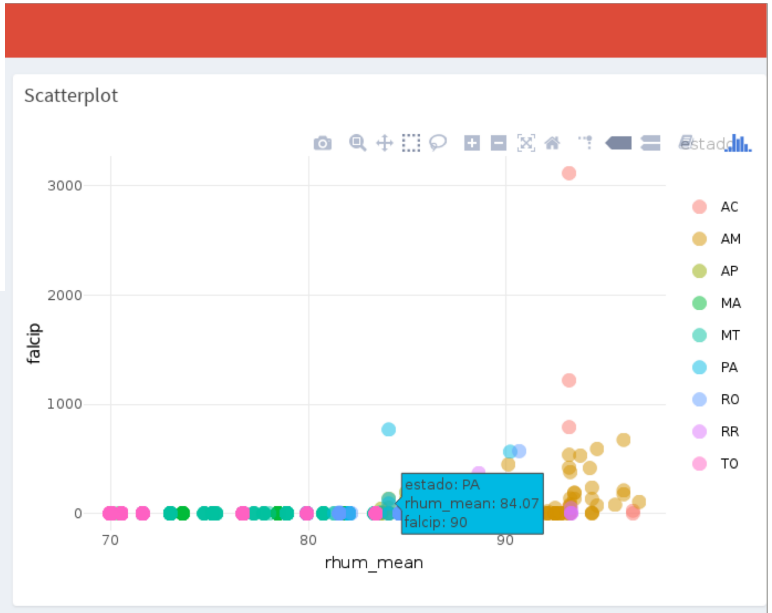
- ✓ Spas
- ✓ Records: 45
- ✓ Attributes: 19

- ✓ Spraying zones
- ✓ Records: 80
- ✓ Attributes: 15



http://200.128.60.86:3838/shiny_integracao/

Descriptive analytics



Descriptive analytics

Juracy Bertoldo

Alberto Sironi

Malaria analytics

- Ecological study (municipalities), spatial and temporal factors
- Classification of malaria epidemics: recurrence and patterns
- Epidemic forecasting model with different granularity
- Inclusion of climate data to support geostatistical analyzes

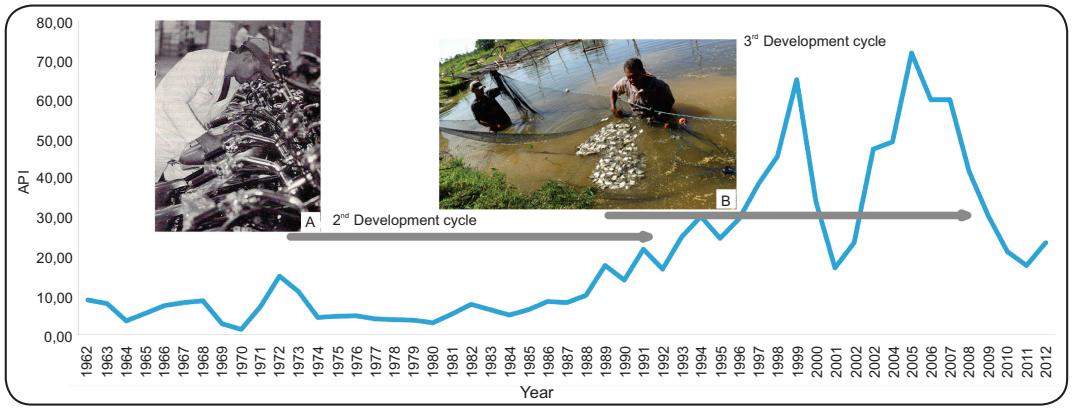


FIGURE 3 - Historical series of malaria cases (1962 to 2012) and contributing factors for disease spread in the Amazon. A: Factory

Review Article

Revista da Sociedade Brasileira de Medicina Tropical 48(Suppl I):4-11, 2015
http://dx.doi.org/10.1590/0037-8682-0275-2014

Malaria in the State of Amazonas: a typical Brazilian tropical disease influenced by waves of economic development

Vanderson Souza Sampaio[1],[2],[3], André Machado Siqueira[4], Maria das Graças Costa Alecrim[1],[2], Maria Paula Gomes Mourão[1],[2], Paola Barbosa Marchesini[5], Bernardino Cláudio Albuquerque[3], Joabi Nascimento[1],[2], Élder Augusto Guimarães Figueira[3], Wilson Duarte Alecrim[1], Wuelton Marcelo Monteiro[1],[2] and Marcus Vinícius Guimarães Lacerda[1],[2],[6]

RESEARCH Open Access

Deforestation, drainage network, indigenous status, and geographical differences of malaria in the State of Amazonas

Wagner Cosme Morhy Terrazas[1], Vanderson de Souza Sampaio[1], Daniel Barros de Castro[1,2], Rosemary Costa Pinto[1], Bernardino Cláudio de Albuquerque[1], Megumi Sadahiro[1], Ricardo Augusto dos Passos[1] and José Ueleres Braga[2,3,4*]

REVIEW Open Access

Malaria in Brazil, Colombia, Peru and Venezuela: current challenges in malaria control and elimination

Judith Recht[1*], André M. Siqueira[2], Wuelton M. Monteiro[3], Sonia M. Herrera[4], Sócrates Herrera[4] and Marcus V. G. Lacerda[3,5]

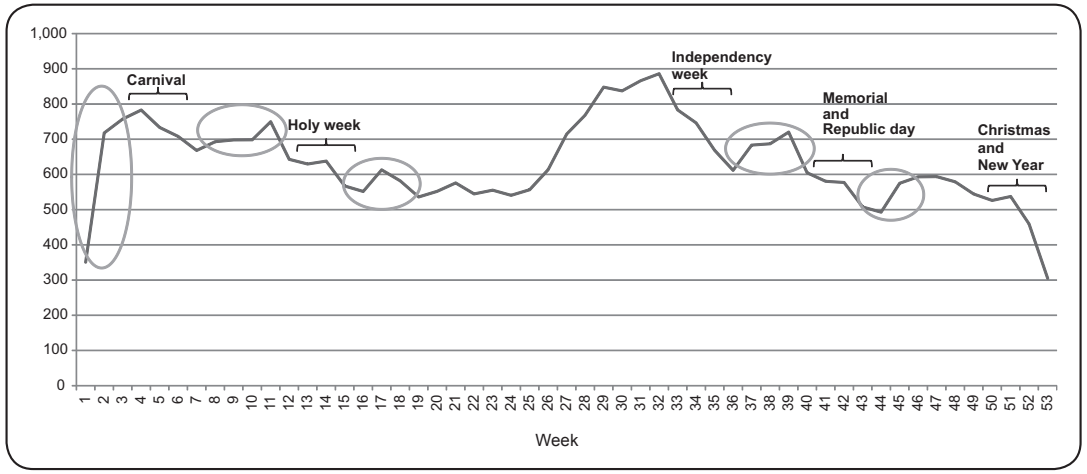
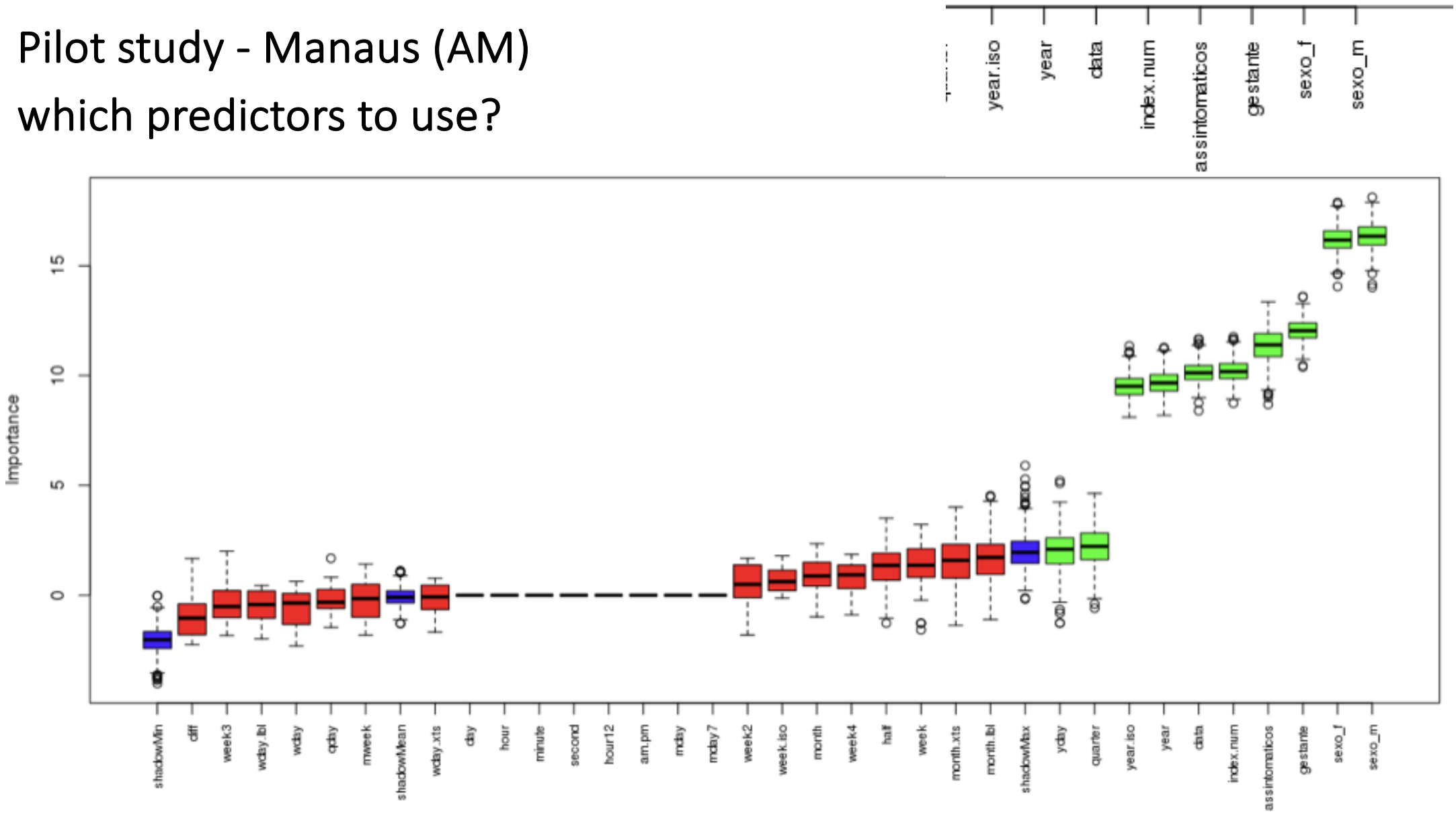


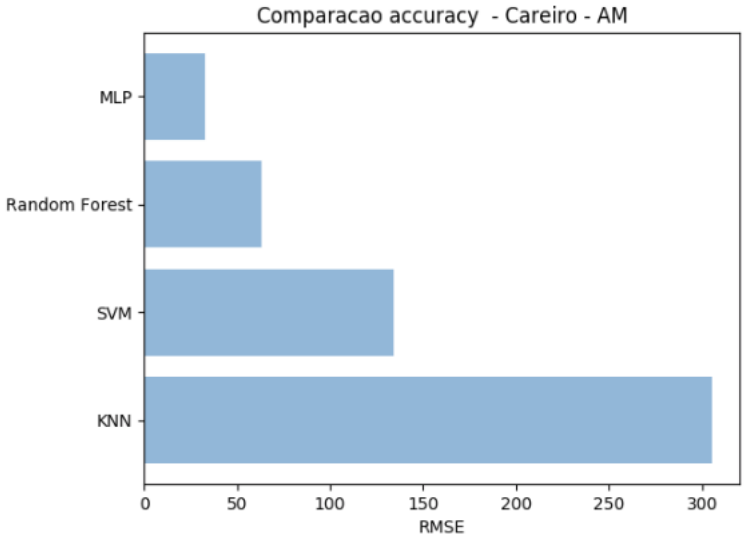
FIGURE 4 - Average annual malaria cases by epidemiological week in Manaus between 2003 and 2012, showing the infection peaks associated with the holidays.

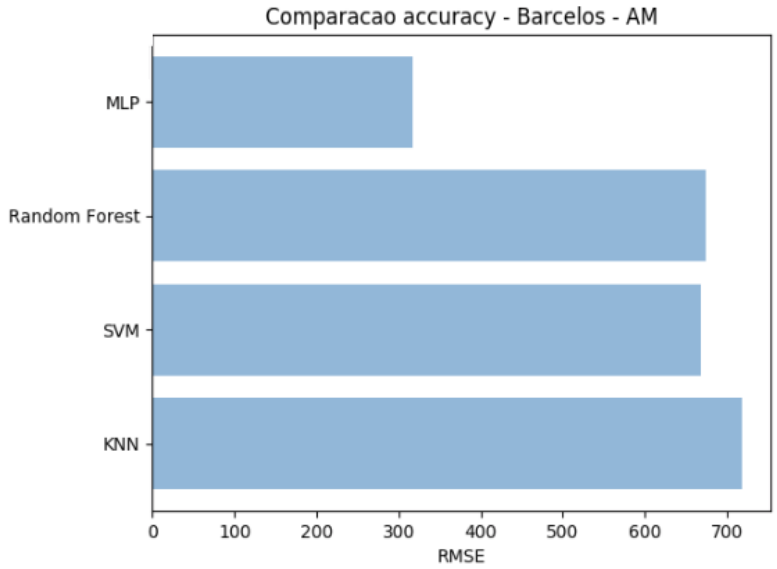
Pilot study - Manaus (AM)

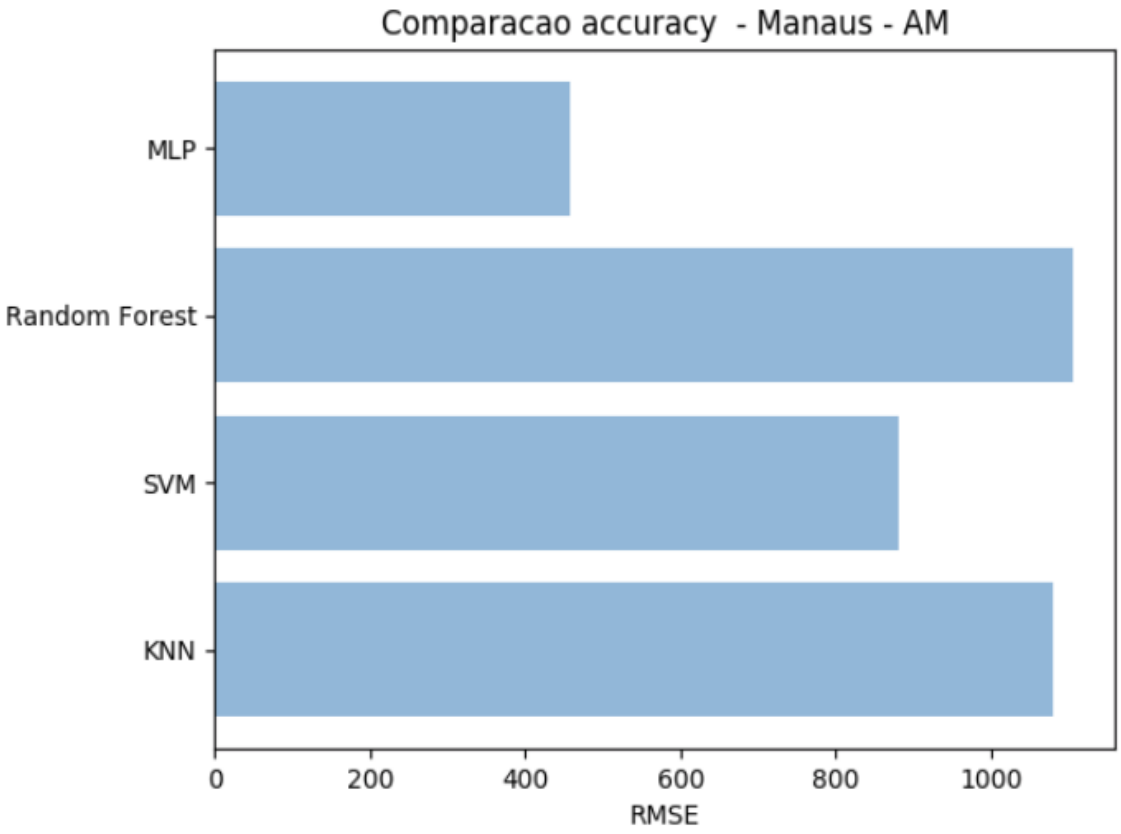
which predictors to use?



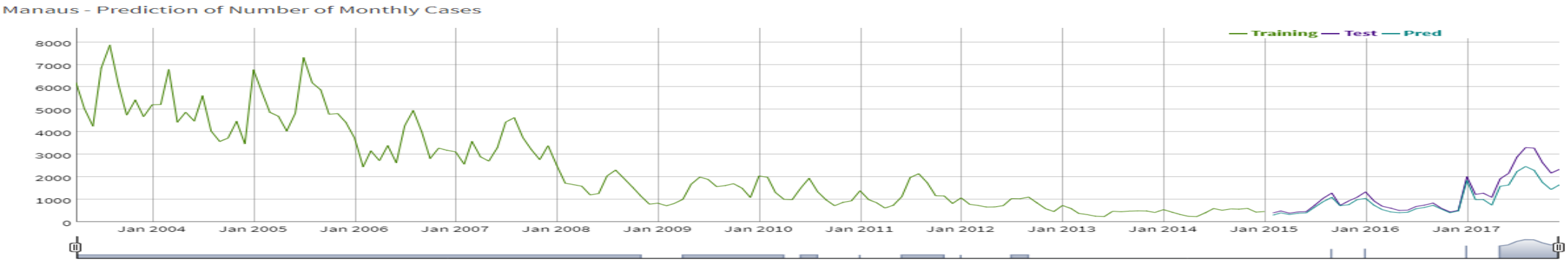
Pilot study - Manaus (AM)

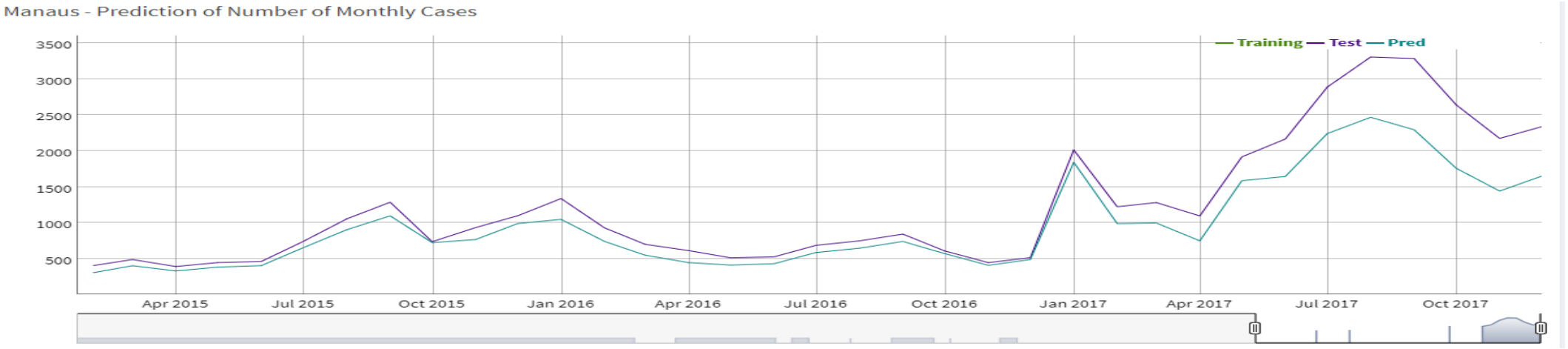






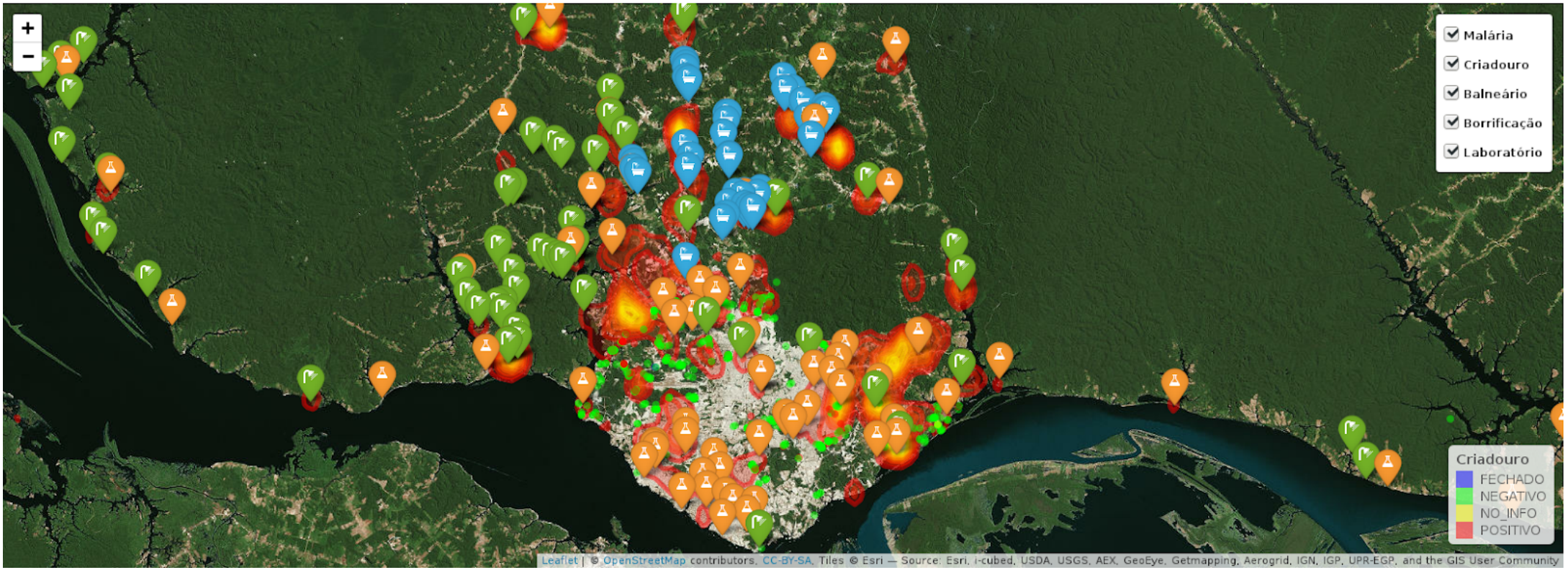
Pilot study - Manaus (AM)





Example: multilayer visual mining

Manaus, capital city of Amazonas



Early Childhood Development Friendly Index

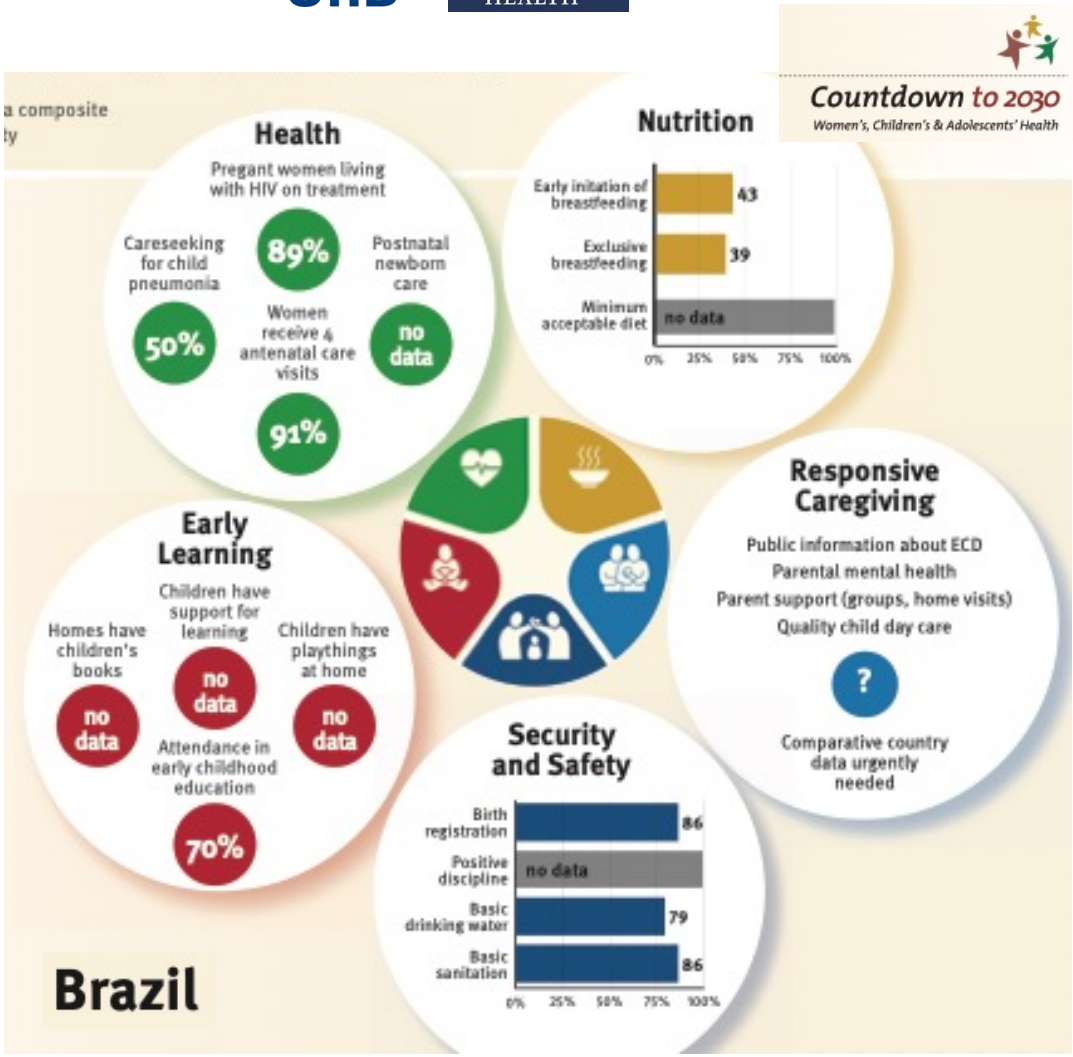
Assessing the enabling environment for Nurturing Care

UnB

Yale SCHOOL OF PUBLIC HEALTH

BILL & MELINDA GATES foundation

Evidence-based nurturing care indicators to assess the factors contributing to enabling environments and promote ECD at the municipal level.



Standardisation of wearable-based algorithms

for healthcare applications in developing countries

Northumbria University NEWCASTLE

ROYAL ACADEMY OF ENGINEERING

npj | Digital Medicine

www.nature.com/npjdigitalmed

PERSPECTIVE OPEN

Not all sensors are created equal: a framework for evaluating human performance measurement technologies

Brian Caulfield[1,2], Brenda Reginatto[2] and Patrick Slevin[3]

Table 1. Three primary application contexts for human performance devices

Application	Use cases
Wellness/Fitness	Personal health/wellness use cases, where the goal is to use data from the device to help a person to better manage their lifestyle.
	Fitness and performance use cases, where the goal is to provide data than can help to guide a training programme for sporting activity.
Healthcare	Behaviour modification use cases, where the goal is to provide input to a structured treatment programme for management of a healthcare issue, or engage patients in their own care process.
	Clinical decision-making process use cases, where the goal is to provide data that can guide diagnosis, treatment decisions or measure outcomes of care.
Clinical trials/Research	Behaviour modification use cases, where the goal is to provide input to a self-directed intervention that might compliment or enhance the impact of a therapeutic product.
	Endpoints, where the goal is to document the impact of a therapeutic product.

Standardisation of wearable-based algorithms

for healthcare applications in developing countries

- ✓ A novel standardised framework to better inform algorithms for a more harmonised gait assessment in Parkinson's disease (PD).
- ✓ Design of an online simulation to test algorithms.
- ✓ Additionally, it will outline an educational process for all clinicians to better understand the functionality of wearables/algorithms and resulting outcomes.

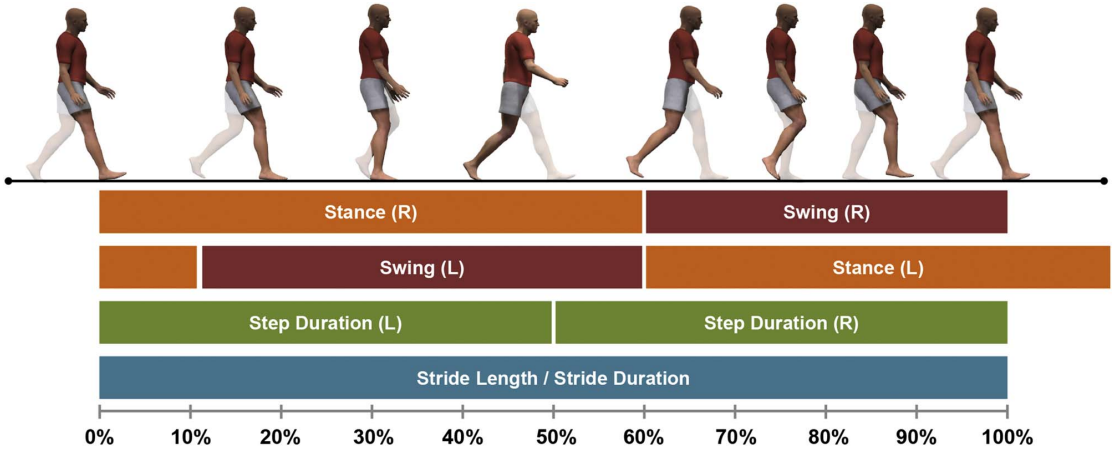
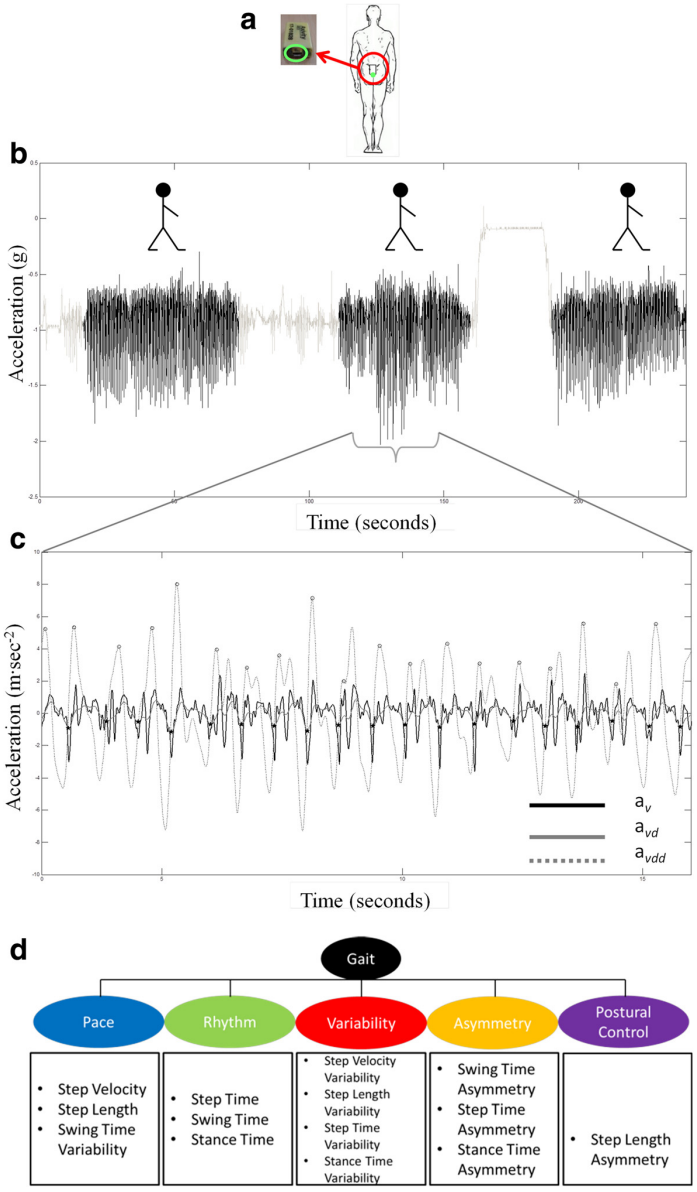


Fig. 1. Schematic of the human gait cycle and the spatiotemporal parameters validated in this study. Specifically, we validated the IMU system's ability to measure the stance percent, swing percent, stride duration (gait cycle time), stride length, and step duration in addition to the speed and cadence of the cycle.



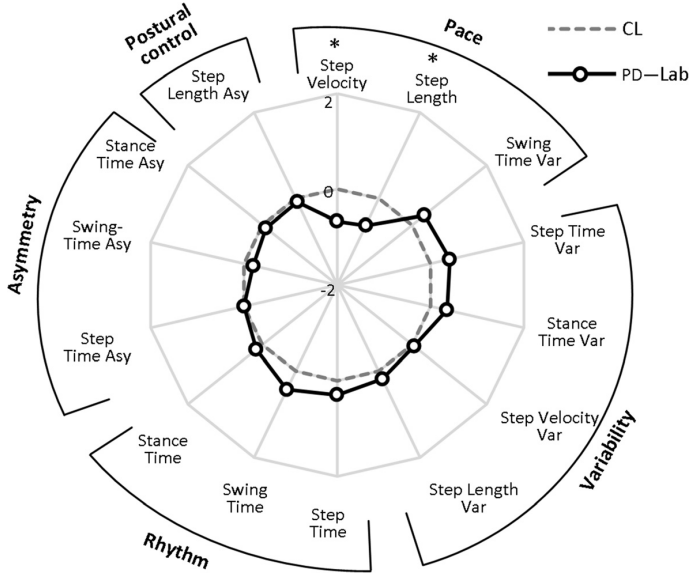


Fig. 3 Radar plot illustrating the 14 gait characteristics organised by domain for people with Parkinson's disease (PD) and controls (CL) evaluated in the laboratory (Lab). The central dotted line represents CL data, deviation from zero along the axis radiating from the centre of the plot represents how many standard deviations (range: ±2 SD, *z* score based on control means and standard deviations) the PD differ from CL. Asterisks represent significant differences between PD and CL (*p* values < 0.01)

Outcome	Method
1. Step detection (step time, stance, swing time)	1. McCamley at al. 2012 (Initial contacts (ICs) and final contacts detection (FCs))
2. Step length	2. Zijlstra et al. 2003 (Inverted pendulum model using ICs evaluated in 1.)
3. Step Velocity	3. Step Velocity = Step Length/ Step Time
4. Variability	4. Standard Deviation (steps)
5. Asymmetry	6. $\lvert Average_{Left}\ (steps) - Average_{Right}\ (steps) \rvert$

Del Din *et al. Journal of NeuroEngineering and Rehabilitation* (2016) 13:46
DOI 10.1186/s12984-016-0154-5

Journal of NeuroEngineering and Rehabilitation

RESEARCH Open Access

Free-living gait characteristics in ageing and Parkinson's disease: impact of environment and ambulatory bout length

Silvia Del Din*, Alan Godfrey, Brook Galna, Sue Lord and Lynn Rochester

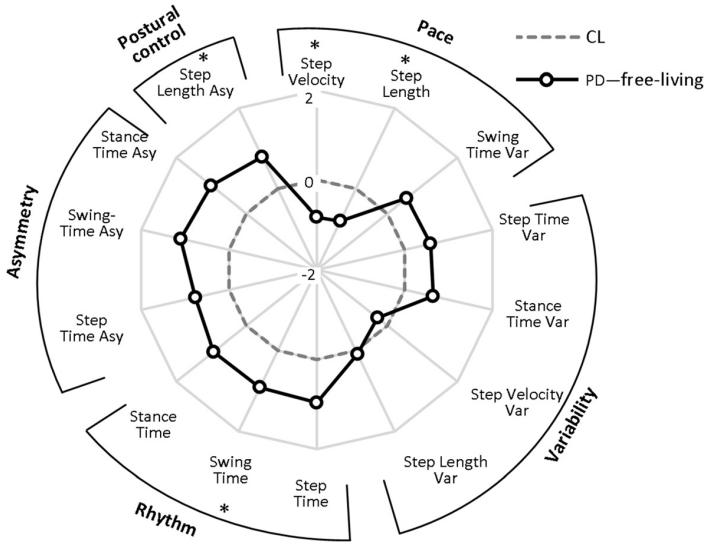


Fig. 4 Radar plot illustrating the 14 gait characteristics organised by domain for people with Parkinson's disease (PD) and controls (CL) evaluated in free-living conditions. The central *dotted line* represents CL data, deviation from zero along the axis radiating from the centre of the plot represents how many standard deviations (range: ± 2 SD, *z* score based on control means and standard deviations) the PD differ from CL. *Asterisks* represent significant differences between PD and CL (*p* values < 0.01)

- **Stratification of patients suffering from myalgic encephalomyelitis/chronic fatigue syndrome.**
  - Principal investigator: Marcos Barreto (UFBA).
  - Team: Nuno Sepulveda (London School of Hygiene & Tropical Medicine), Robespierre Pita (UFBA)
  - Period: 2019-2020
  - Scope: This study aims at to stratify ME/CFS patients into different clusters (or symptom subtypes). The respective objectives are the following: i) to distinguish ME/CFS patients from those suffering from multiple sclerosis (MS); ii) to identify sets of clinical symptoms that could characterize different clusters of ME/CFS patients; iii) to identify the best (or exclusive) predictive symptoms for CFS and compare the results with those obtained from different statistical/computational methods; (iv) to compare the stability of patients stratification using baseline and follow-up data.

Support:

LONDON SCHOOL of HYGIENE &TROPICAL MEDICINE

CUREME

biobank uk
Improving the health of future generations

ICE FALCON – Inference on causation from examination of familial confounding

- ✓ multivariate twin model applied to opposite-sex twin data.
- ✓ simulation study, comparison with Mendelian Randomisation.
- ✓ apply to continuous, binary (Pearson) and time-to-event (Cox) outcomes.
- ✓ develop zygosity prediction algorithms.

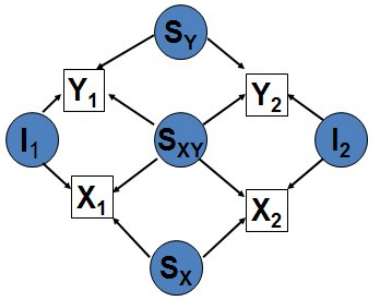


Figure 2

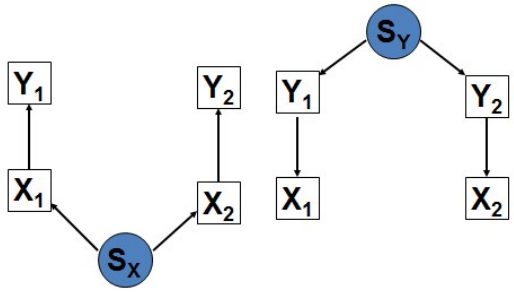


Figure 3

In terms of a regression equation (Gurrin et al., 2006),
$Y_i = b_{self}\, X_i + b_{co\text{-}twin}\, X_j + E_i$, for i, j = 1,2 $i \neq j$, (E_i is error).
We fit Model I: $Y_i = b_{self}\, X_i + E_i$; Model II: $Y_i = b_{co\text{-}twin}\, X_j + E_i$; and Model III: $Y_i = b_{self}\, X_i + b_{co\text{-}twin}\, X_j + E_i$.

Table 1. Expected results when fitting X as a predictor of Y for different causal scenarios

Model	Coeff	Familial confounding	X causes Y	Y causes X
I	b_{self}	Marginal association	Marginal association	Marginal association
II	$b_{co\text{-}twin}$	Marginal association	Marginal association	No marginal association
III	adjusted b_{self}	Association attenuated compared with b_{self}	Association the same as b_{self} of Model 1	Association the same as b_{self} of Model 1
III	adjusted $b_{co\text{-}twin}$	Association attenuated compared with $b_{co\text{-}twin}$	Null (attenuated compared with $b_{co\text{-}twin}$)	Association

Brazilian Journal of Physical Therapy 2018;22(3):184–189

abrapg-ft Associação Brasileira de Pesquisa e Pós-Graduação em Fisioterapia

Brazilian Journal of Physical Therapy

https://www.journals.elsevier.com/brazilian-journal-of-physical-therapy

BJPT

MASTERCLASS

Twin studies for the prognosis, prevention and treatment of musculoskeletal conditions

Lucas Calais-Ferreira[a,*], Vinicius C. Oliveira[b], Jeffrey M. Craig[c,d,e], Louisa B. Flander[a], John L. Hopper[a], Luci F. Teixeira-Salmela[f], Paulo H. Ferreira[g]

Arthritis Care & Research

American College of Rheumatology Empowering Rheumatology Professionals

Hallux Valgus | Free Access

Hallux Valgus, By Nature or Nurture? A Twin Study

Shannon E. Munteanu, Hylton B. Menz, John D. Wark, Jemma J. Christie, Katrina J. Scurrah, Minh Bui, Bircan Erbas, John L. Hopper, Anita E. Wluka

First published: 18 November 2016 | https://doi.org/10.1002/acr.23154 | Cited by: 6

Psychological Medicine

Article | Supplementary materials | Metrics | Correction

Volume 46, Issue 15 November 2016 , pp. 3213-3218

The effects of stress–tension on depression and anxiety symptoms: evidence from a novel twin modelling analysis

C. G. Davey (a1) (a2) (a3), C. López-Solà (a4) (a5), M. Bui (a6), J. L. Hopper (a6) (a7) ...

https://doi.org/10.1017/S0033291716001884 Published online: 08 September 2016

THE UNIVERSITY OF MELBOURNE

Australian Government
National Health and Medical Research Council

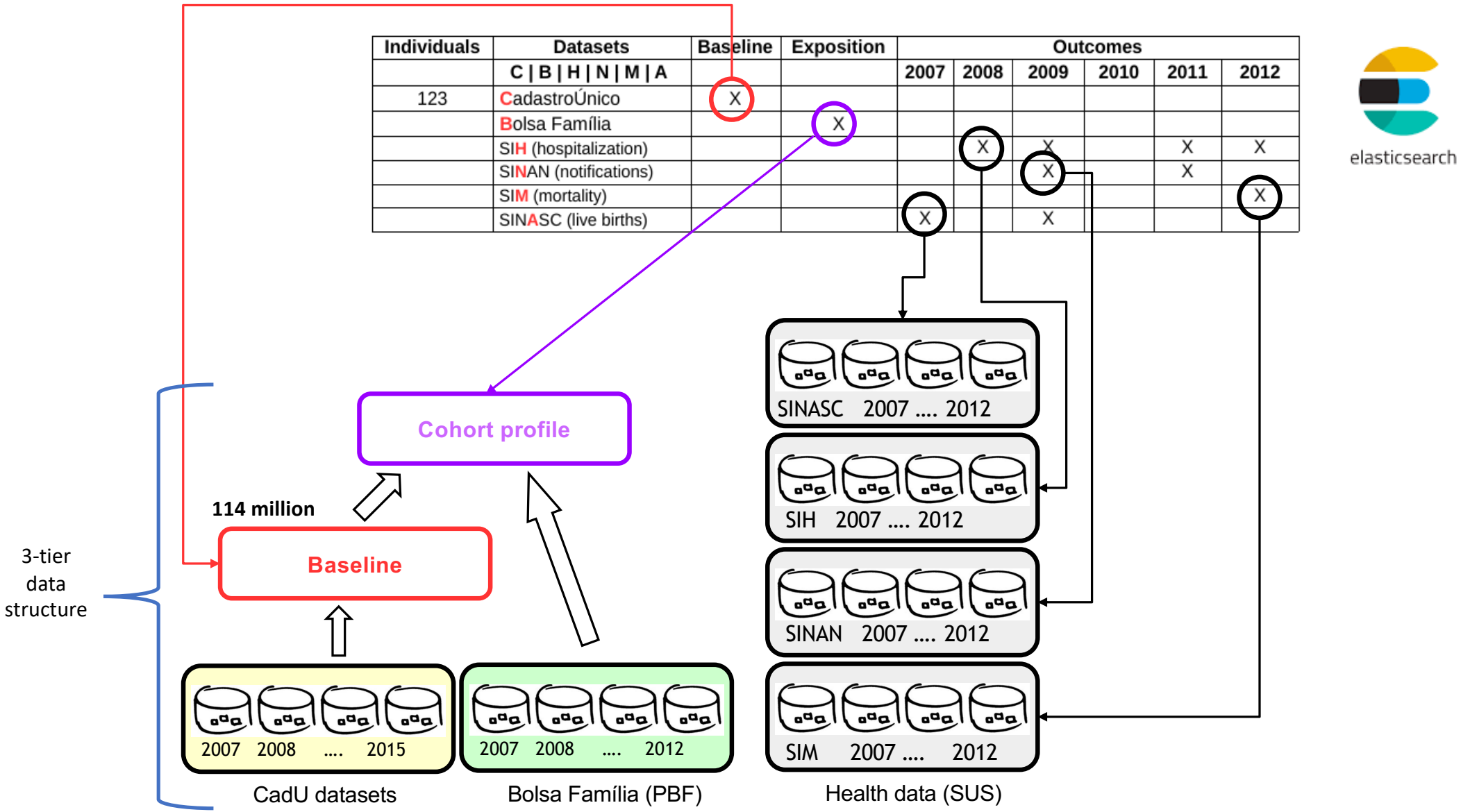
OBRIGADO!
GRACIAS!
THANK YOU!

www.atyimolab.ufba.br

Dr. Marcos Barreto

marcosb@ufba.br

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			



elasticsearch

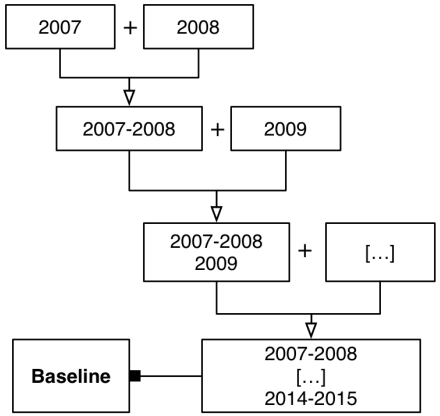
CADU 114 million records cohort

Year	Table	File size (GB)	# of records
2007	A	11,4	21,028,364
	B	86,8	79,050,446
2008	A	12,5	22,767,472
	B	100,1	89,915,568
2009	A	13,5	24,661,693
	B	108,8	97,640,845
2010	A	14,3	26,107,223
	B	114,4	102,663,287
2011	1	25,0	27,014,194
	4	4,3	106,433,938
2012	1	11,0	30,268,867
	4	27,0	115,636,503
2013	1	6,5	32,897,120
	4	29,0	123,116,446
2014	1	7,1	35,439,015
	4	34,0	130,430,300
2015	1	7,6	35,912,231
	4	36,0	136,368,326

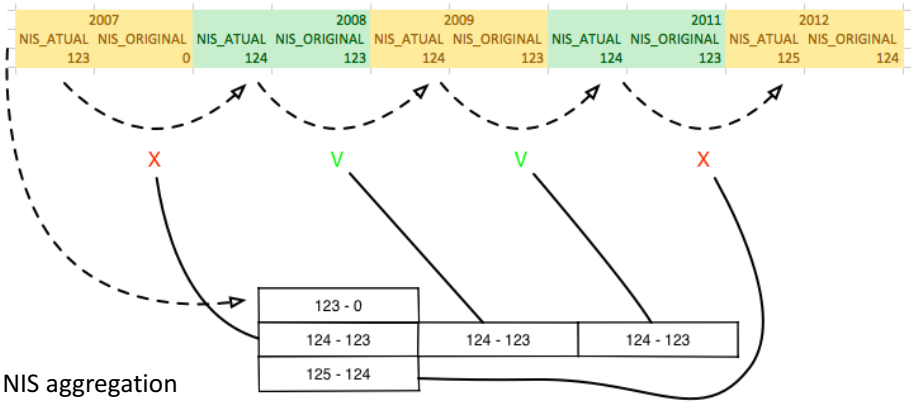
CADU v6 (2007-2010), 2 tables, 167 attributes
CADU v7 (2011 onwards), 18 tables, 433 attributes

+ sensitive data

Progressive merge



15 attributes



NIS aggregation

Versão 6	Versão 7	Descrição das variáveis
NM_PESSOA	NOM_PESSOA	Nome da pessoa
CD_FAMILIAR	COD_FAMILIAR_FAM	Código que identifica a família da pessoa, contém 11 dígitos
NU_NIS_PESSOA	NUM_NIS_PESSOA_ATUAL	Número do NIS (Número de identificação Social) atual da pessoa, contém
DT_NASCIMENTO	DTA_NASC_PESSOA	Data de nascimento da pessoa, no forma
CD_SEXO	COD_SEXO_PESSOA	Código que indica o sexo da pessoa
CD_PARENTESCO	COD_PARENTESCO_RF_PESSOA	Código dos parentescos da pessoa
CD_RACA_COR	COD_RACA_COR_PESSOA	Código da raça da pessoa
NM_MAE	NOM_COMPLETO_MAE_PESSOA	Nome completo da mãe da pessoa
NM_PAI	NOM_COMPLETO_PAI_PESSOA	Nome completo do pai da pessoa
CD_IBGE_NASCIMENTO	COD_IBGE_MUNIC_NASC_PESSOA	Código do municipio de nascimento da pessoa
	COD_MUNIC_IBGE_2_FAM	Código que indica o estado em que a pessoa mora
CD_IBGE_LOGRADOURO	COD_MUNIC_IBGE_5_FAM	Código que indica o municipio que a pessoa mora
CD_PAIS_ORIGEM	COD_PAIS_ORIGEM_PESSOA	Código do país de origem dos pais das pessoas
NU_CPF	NUM_CPF_PESSOA	Número do CPF (Cadastro de Pessoa física) da pessoa
NU_DOCUMENTO_IDENTIDADE	NUM_IDENTIDADE_PESSOA	Número do RG(Registro Geral) da pessoa
NU_CTPS	NUM_CART_TRAB_PREV_SOC_PESSOA	Número do CTPS(Carteira de Trabalho e Previdência Social) da pessoa
NU_TITULO_ELEITOR	NUM_TITULO_ELEITOR_PESSOA	Número do tíutulo de eleitor na pessoa
DT_PESQUISA_CADASTRO	DAT_CADASTRAMENTO_FAM	Data de cadastramento da família
IN_DEFICIENCIA_CEGUEIRA	IND_DEF_CEGUEIRA_MEMB	Indica se o membro familiar apresenta cegueira
IN_DEFICIENCIA_SURDEZ	IND_DEF_SURDEZ_PROFUNDA_MEMB	Indica se o membro familiar apresenta surdez profunda
	ND_DEF_SURDEZ_LEVE_MEMB	Indica se o membro familiar apresenta surdez leve
IN_DEFICIENCIA_FISICA	IND_DEF_FISICA_MEMB	Indica se o membro familiar apresenta deficiência física
IN_DEFICIENCIA_MENTAL	IND_DEF_MENTAL_MEMB	Indica se o membro familiar apresenta mental
CD_ESCOLA	IND_FREQUENTA_ESCOLA_MEMB	Indica o tipo de instituição escolar que o membro frequentou
CD_MERCADO_TRABALHO	COD_PRINCIPAL_TRAB_MEMB	Código do principal trabalho do membro
VL_REMUNERACAO_EMPREGO	VAL_REMUNER_EMPREGO_MEMB	Valor da remuneração do emprego do membro
VL_RENDA_APOSENTADORIA	VAL_RENDA_APOSENT_MEMB	Valor da renda do membro aposentado
VL_RENDA_SEGURO_DESEMPREGO	VAL_RENDA_SEGURO_DESEMP_MEMB	Valor da renda do seguro desemprego recebido do membro
VL_RENDA_PENSAO_ALIMENTICIA	VAL_RENDA_PENSAO_ALIMEN_MEMB	Valor da renda recebida por pensão alimenticia do membro
VL_OUTRAS_RENDAS	VAL_OUTRAS_RENDAS_MEMB	Valor de outras rendas do membro
QT_PESSOAS_INFORMADA	QTD_PESSOAS_DOMIC_FAM	Quantidade de pessoas no domicílio familiar
NU_COMODOS	QT_COMODOS_DOMIC_FAM	Quantidade de comodos no domicílio familiar
CD_TIPO_LOCALIDADE	COD_LOCAL_DOMIC_FAM	Código do local do domicílio familiar
CD_SITUACAO_DOMICILIO	COD_ESPECIE_DOMIC_FAM	Código especie do domicílio familiar
CD_CONSTRUCAO	COD_MATERIAL_DOMIC_FAM	Código do material do domicílio familiar
CD_ABASTECIMENTO_AGUA	COD_ABASTE_AGUA_DOMIC_FAM	Código do abastecimento de água do domicílio familiar
CD_ESCOAMENTO_SANITARIO	COD_ESCOA_SANITARIO_DOMIC_FAM	Código do escoamento sanitário do do domicílio familiar
CD_DESTINO_LIXO	COD_DESTINO_LIXO_DOMIC_FAM	Código do destino de lixo do domicílio familiar
CD_ILUMINACAO	COD_ILUMINACAO_DOMIC_FAM	Código da iliminação do domicílio familiar
VL_DESPESA_LUZ	VAL_DESP_ENERGIA_FAM	Valor de despesa com energia na família
VL_DESPESA_AGUA	VAL_DESP_AGUA_ESGOTO_FAM	Valor de despesa com água e esgoto na família
VL_DESPESA_GAS	VAL_DESP_GAS_FAM	Valor de despesa com gás na família
VL_DESPESA_ALIMENTACAO	VAL_DESP_ALIMENTACAO_FAM	Valor de despesa com alimentação na família
VL_DESPESA_TRANSPORTE	VAL_DESP_TRANSPOR_FAM	Valor de despesa com transporte na família
VL_DESPESA_ALUGUEL	VAL_DESP_ALUGUEL_FAM	Valor de despesa com aluguel na família
VL_DESPESA_MEDICAMENTOS	VAL_DESP_MEDICAMENTOS_FAM	Valor de despesa com medicamentos na família
NM_ESTABELECIMENTO_SAUDE	NOM_ESTAB_ASSIST_SAUDE_FAM	Nome do estabelicimento de assistência a saúde da família

CADU cohort profile: personal + socioeconomic data

AtyImo – trainable model to accuracy assessment