CIÊNCIA DE DADOS E SUA APLICAÇÃO EM SAÚDE

Prof. Marcos Barreto

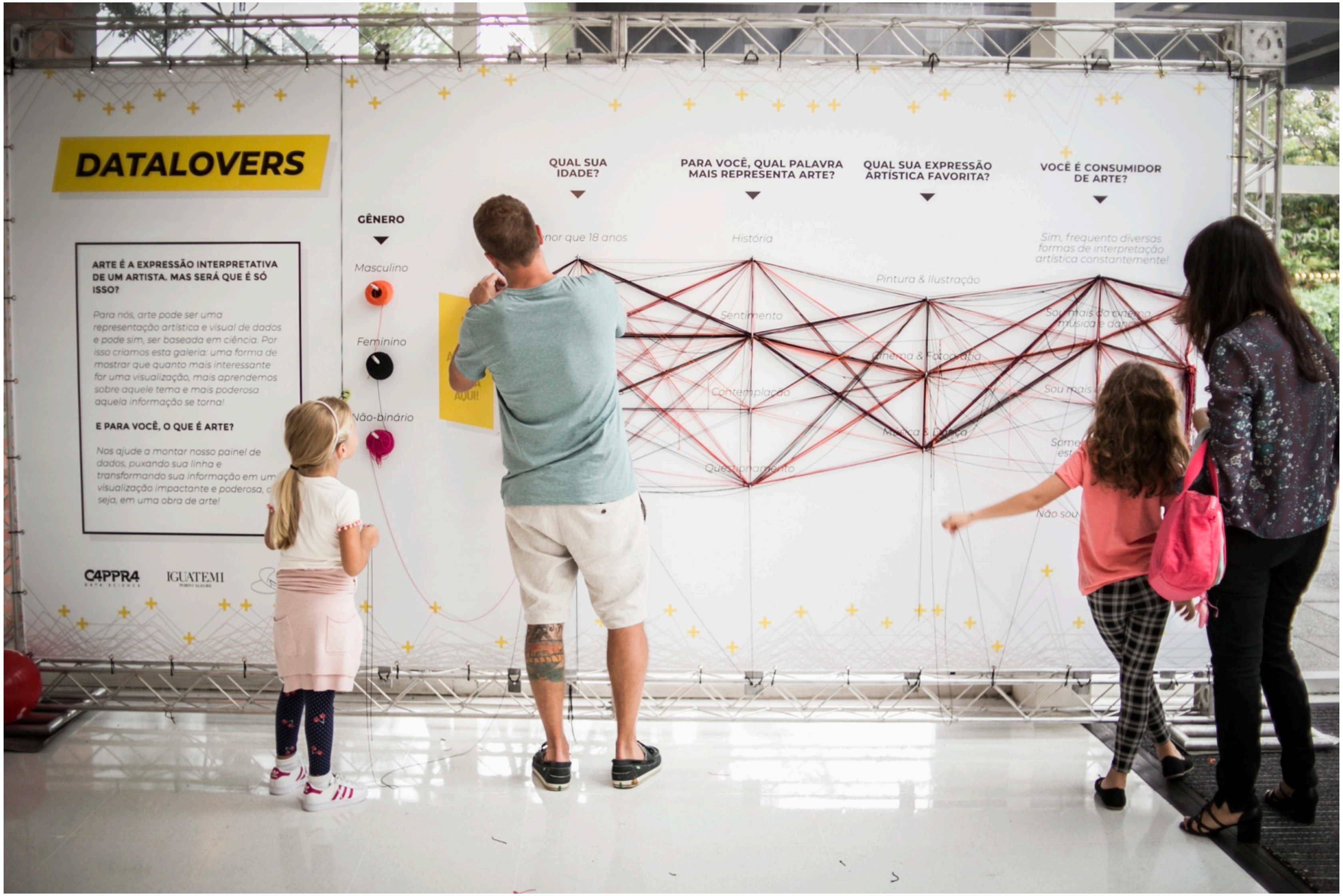
AtyImoLab / LaSiD
Departamento de Ciência da Computação
Instituto de Matemática e Estatística
Universidade Federal da Bahia (UFBA)

Salvador, 22 de março de 2018

Popularização da Ciência de Dados

=

Popularização da Ciência de Dados



https://cappra.com.br/2018/04/26/apaixonados-por-dados-datalovers/

Ciência de dados

[ocultar]

Origem: Wikipédia, a enciclopédia livre.

Ciência de dados (em inglês: *data science*) é uma área interdisciplinar voltada para o estudo e a análise de dados, estruturados ou não, que visa a extração de conhecimento ou *insights* para possíveis tomadas de decisão, de maneira similar à mineração de dados. Ciência de dados alia big data e machine learning, além de técnicas de outras áreas interdisciplinares como estatística, economia, engenharia e outros subcampos da computação como: banco de dados e análise de agrupamentos (*cluster analysis*). A ciência de dados é um campo que já existe há 30 anos, porém ganhou mais destaque nos últimos anos devido a alguns fatores como: o surgimento e popularização do Big Data e o desenvolvimento de áreas como o machine learning. A ciência de dados pode, por exemplo, transformar essa grande quantidade de dados brutos em *insights* de negócios, e com isso, auxiliar empresas em tomadas de decisões para atingir melhores resultados.[1]

Is Data Science a science?

https://towardsdatascience.com/data-science-and-ai-for-business-data-analysts-64f28a5d7ff2

História (base Estatística)

1962 (John Turkey)
The future of data analysis

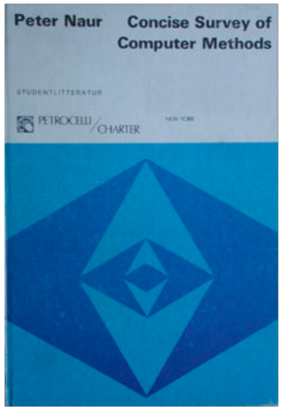
Data analysis, and the parts of statistics which adhere to it, must…take on the characteristics of science rather than those of mathematics… data analysis is intrinsically an empirical science…

1968 (IFIP)
Datalogy

Datalogy, the science of data and of data processes and its place in education…

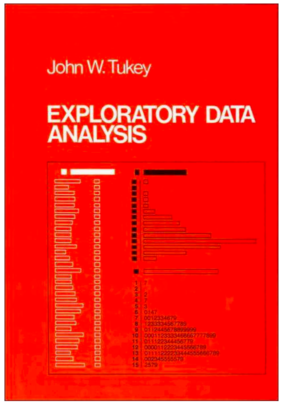
1974 (Peter Naur)
Concise Survey of Computer Methods

“[Data is] a representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process.”



1977 (John Turkey)
Exploratory data analysis

“…more emphasis needed to be placed on using data to suggest hypotheses to test and that Exploratory Data Analysis and Confirmatory Data Analysis should proceed side by side."



1977 (ISI)
International Association for Statistical Computing

“…link traditional statistical methodology, modern computer technology, and the knowledge of domain experts in order to convert data into information and knowledge.”

https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#78699be255cf

História (base Computação + BI)

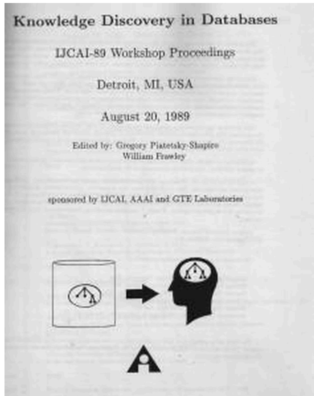
“Companies are collecting mountains of information about you, crunching it to predict how likely you are to buy a product, and using that knowledge to craft a marketing message precisely calibrated to get you to do so…”

The classification societies have variously used the terms data analysis, data mining, and data science in their publications.

“Historically, the notion of finding useful patterns in data has been given a variety of names, including data mining, knowledge extraction, information discovery, information harvesting, data archeology, and data pattern processing..."

1989 (Piatetsky-Shapiro)

1st KDD workshop



1994 (Business Week)

Business

Database Marketing

Jonathan Berry

5 de setembro de 1994 01:00 BRT

1996 (IFCS)

International Federation of Classification Societies

1996 (Fayyad - Shapiro)

From data mining to KDD paper

2001 (William Cleveland)

Data Science Action Plan

It is a plan “to enlarge the major areas of technical work of the field of statistics.”

https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#78699be255cf

História (base Computação + BI)

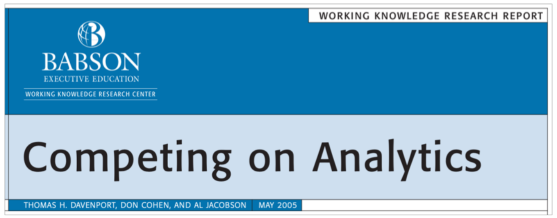
2001 (Breiman)
Statistical modeling: the two cultures

2002 (CODATA)

2003
Journal of Data Science

JDS

2005 (Babson)



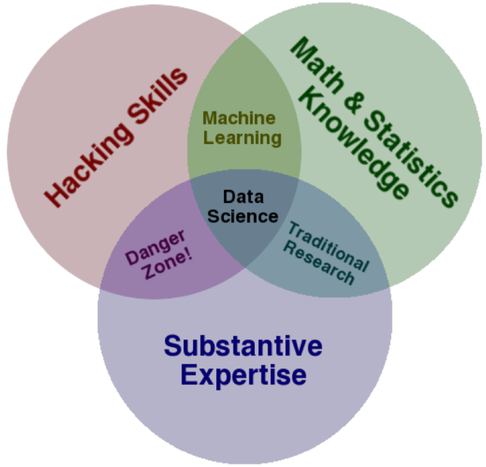
2007 (Zhu)
Introduction to Dataology

2008 - 2009

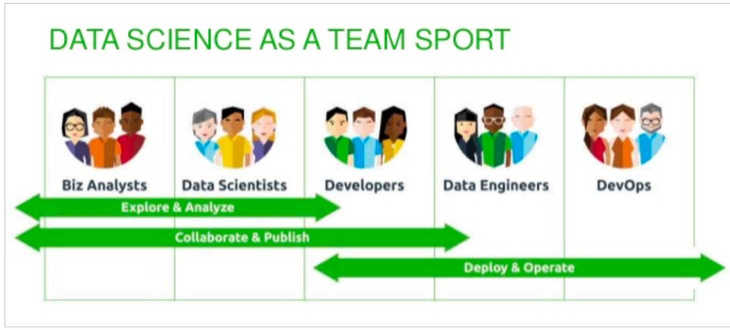
Diversas publicações sobre a necessidade de habilidades específicas para lidar com esta nova área

2010

Diversas "novas" definições para ciência de dados



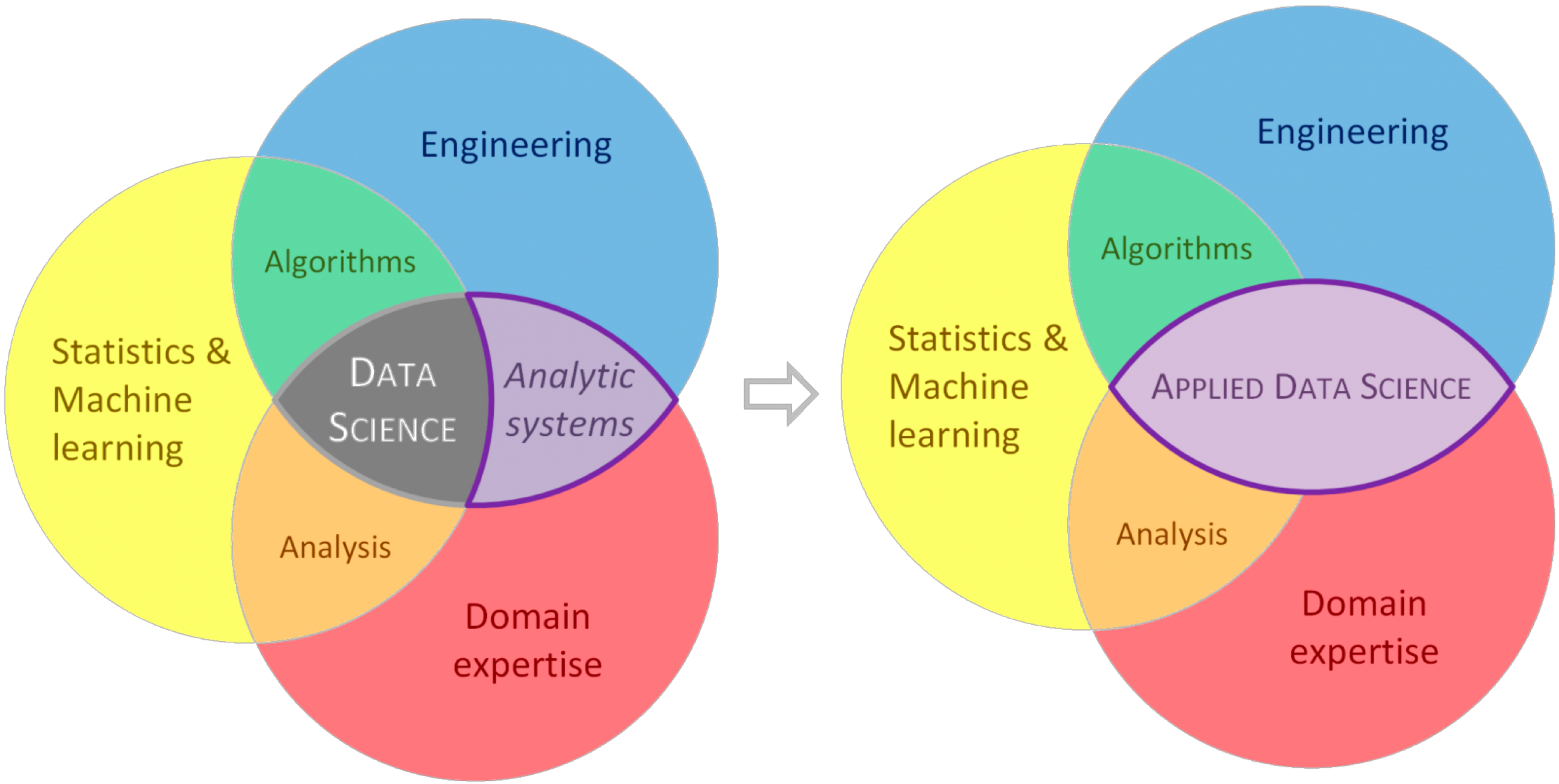
2011



2012 (Davenport)
Data Scientist: The Sexiest Job of the 21st Century

https://www.forbes.com/sites/gilpress/2013/05/28/a-very-short-history-of-data-science/#78699be255cf

Definindo Ciência de Dados

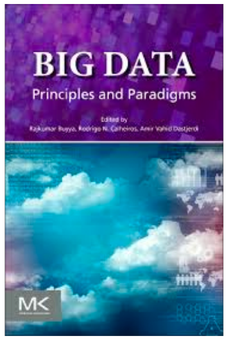


BIG DATA


Big data is like teenage sex:
everyone talks about it,
nobody really knows how to do it,
everyone thinks everyone else is
doing it, so everyone claims they
are doing it...
(Dan Ariely)

Big Data

1. Volume stands for scale of data
2. Velocity denotes the analysis of streaming data
3. Variety indicates different forms of data
4. Veracity focuses on trustworthiness of data sources
5. Variability refers to the complexity of data set. In comparison with "Variety" (or different data format), it means the number of variables in data sets
6. Visibility emphasizes that you need to have a full picture of data in order to make informative decision



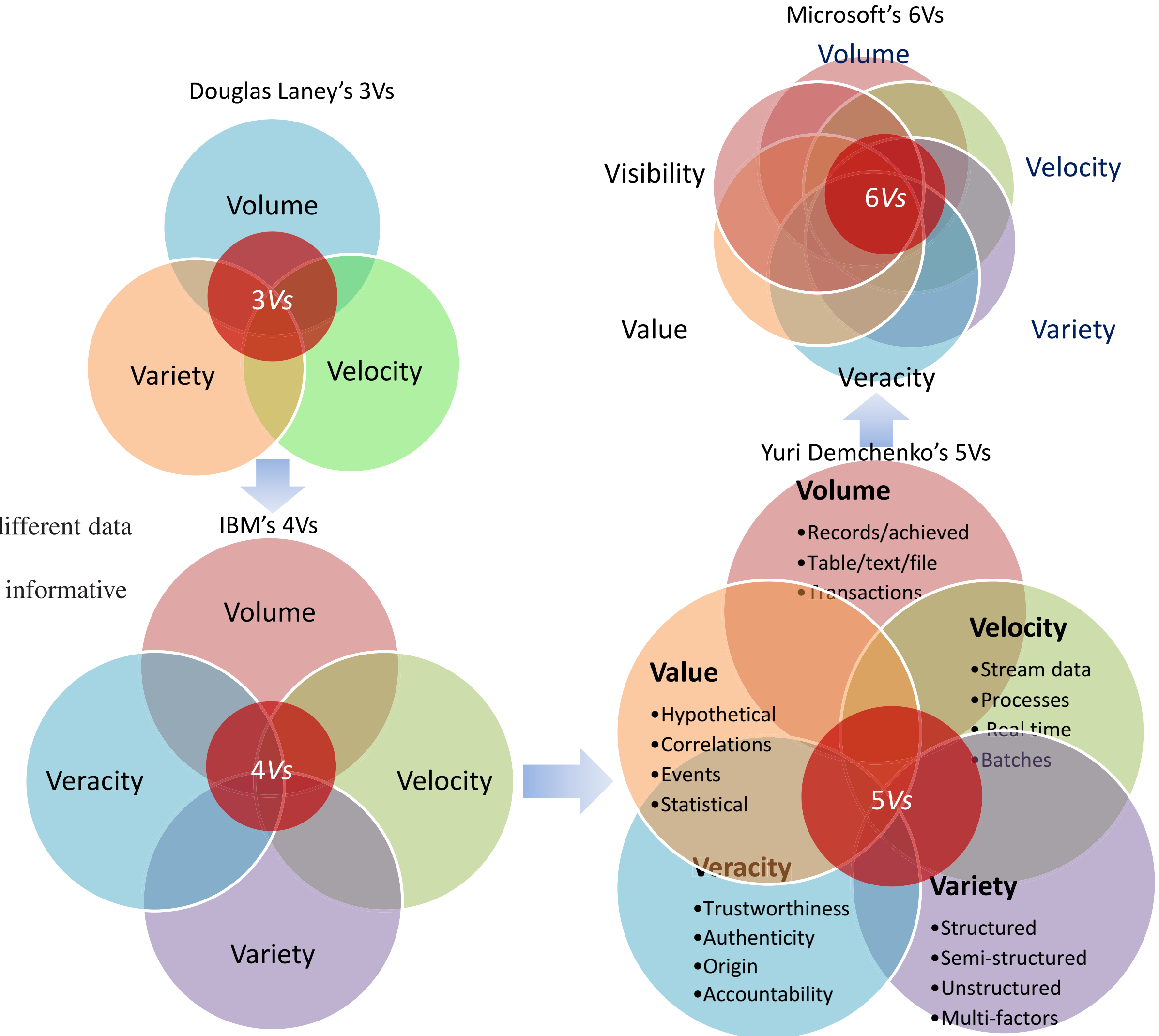


FIG. 3

From 3Vs, 4Vs, 5Vs, and 6Vs big data definition.

Big Data

- Data domain (searching for patterns)
- Business intelligence domain (making predictions)
- Statistical domain (making assumptions)

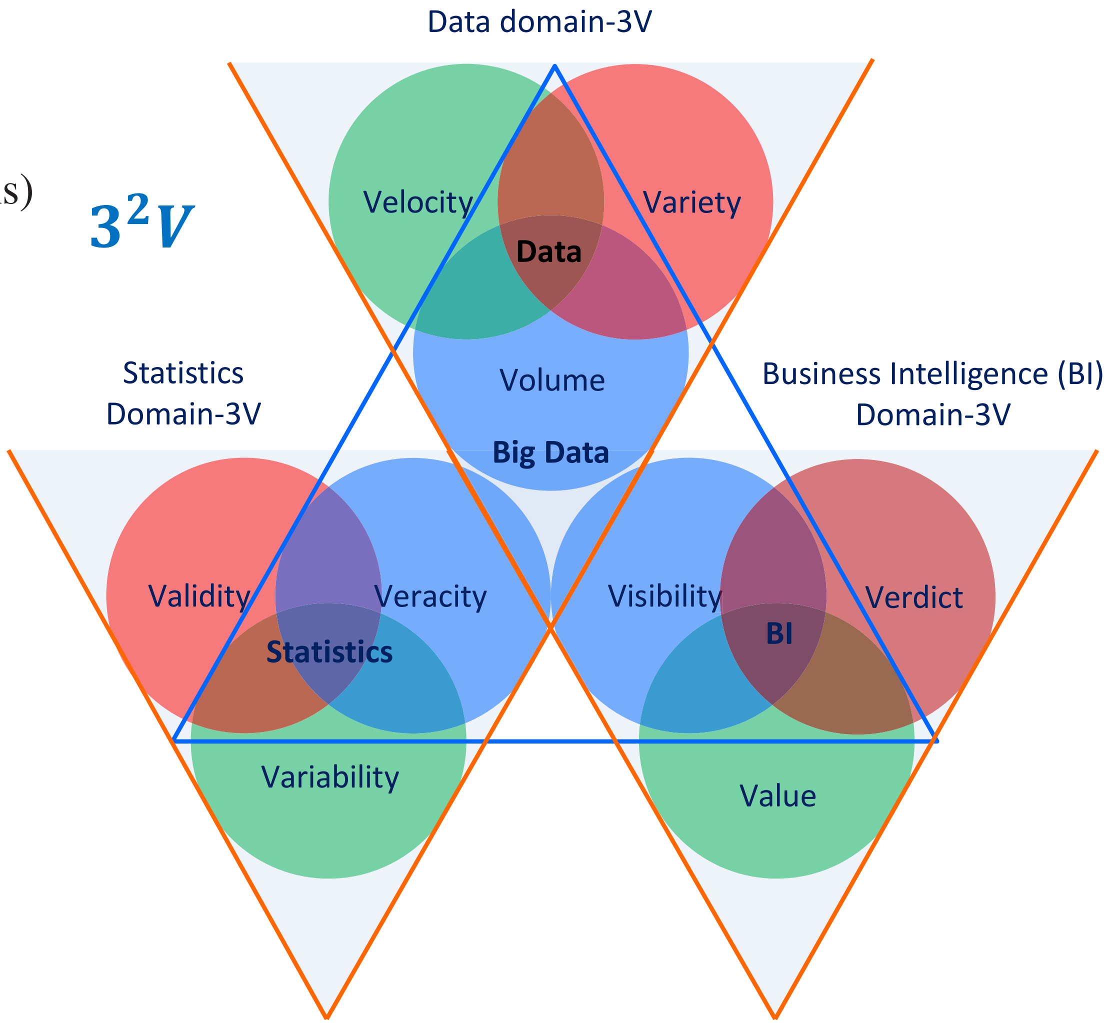
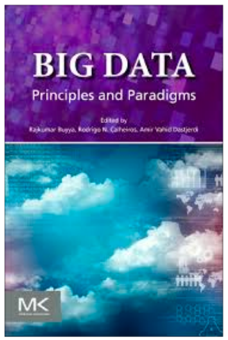
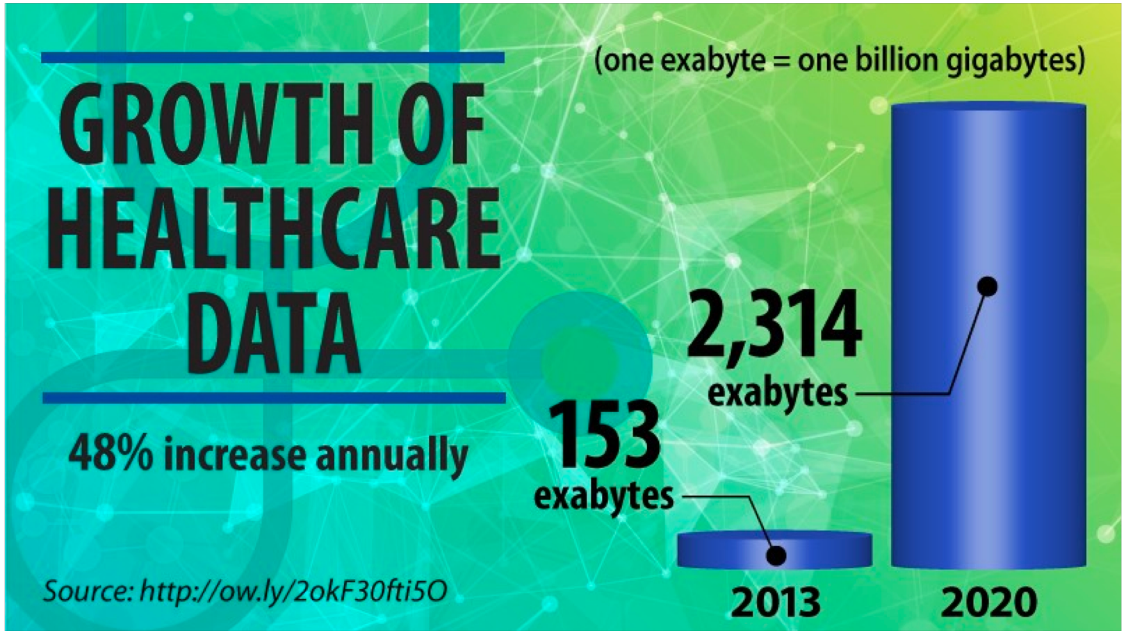


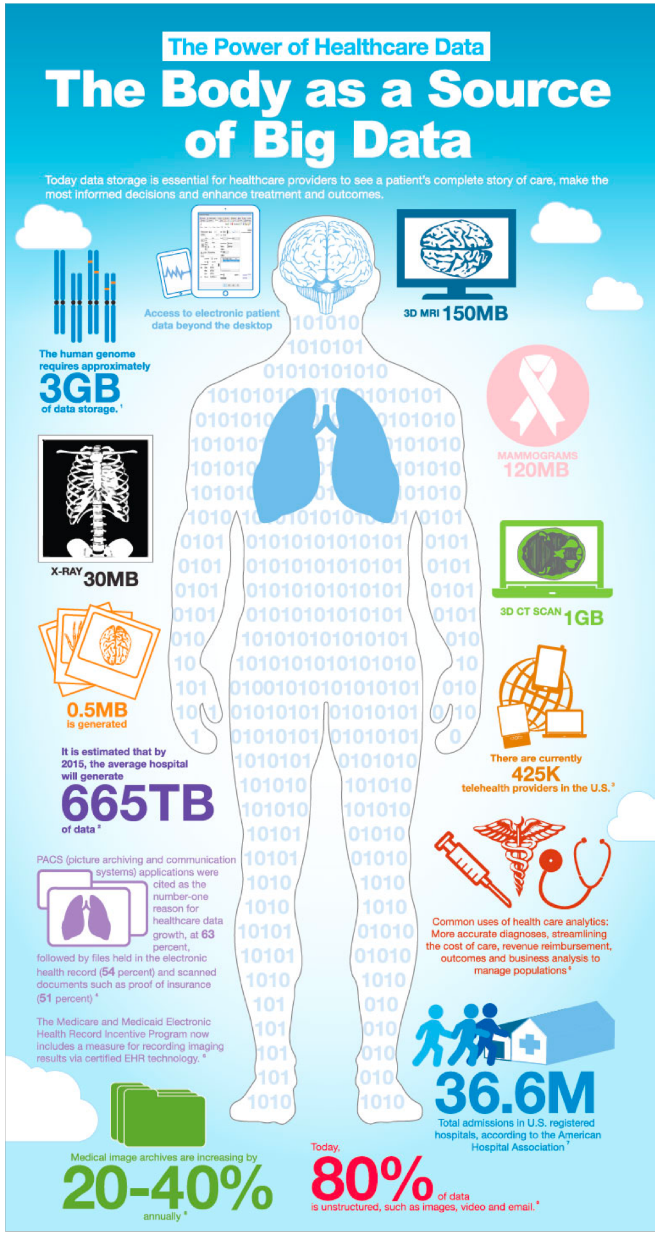
FIG. 5

3^2Vs Venn diagrams in hierarchical model.



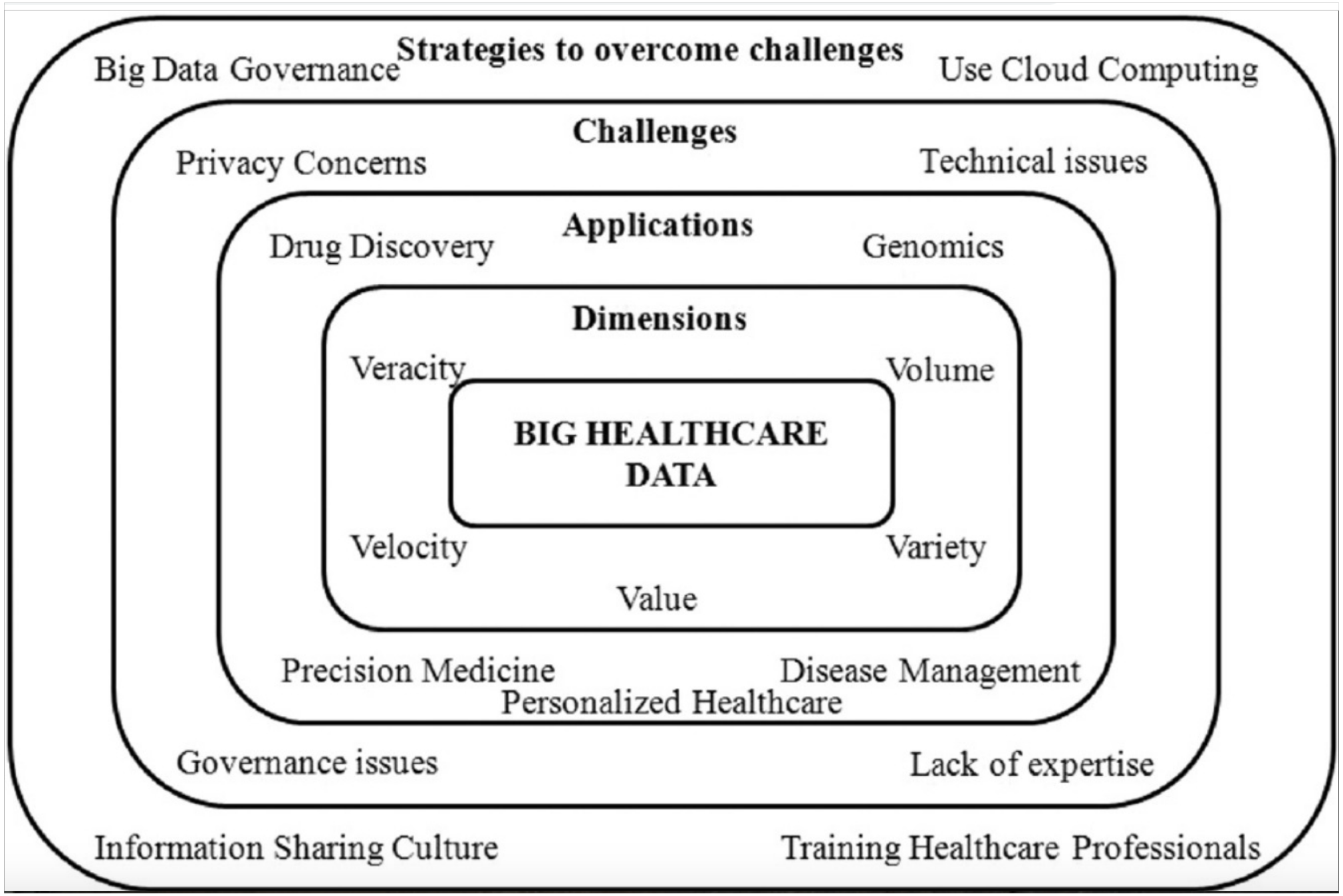
Big Data em Saúde



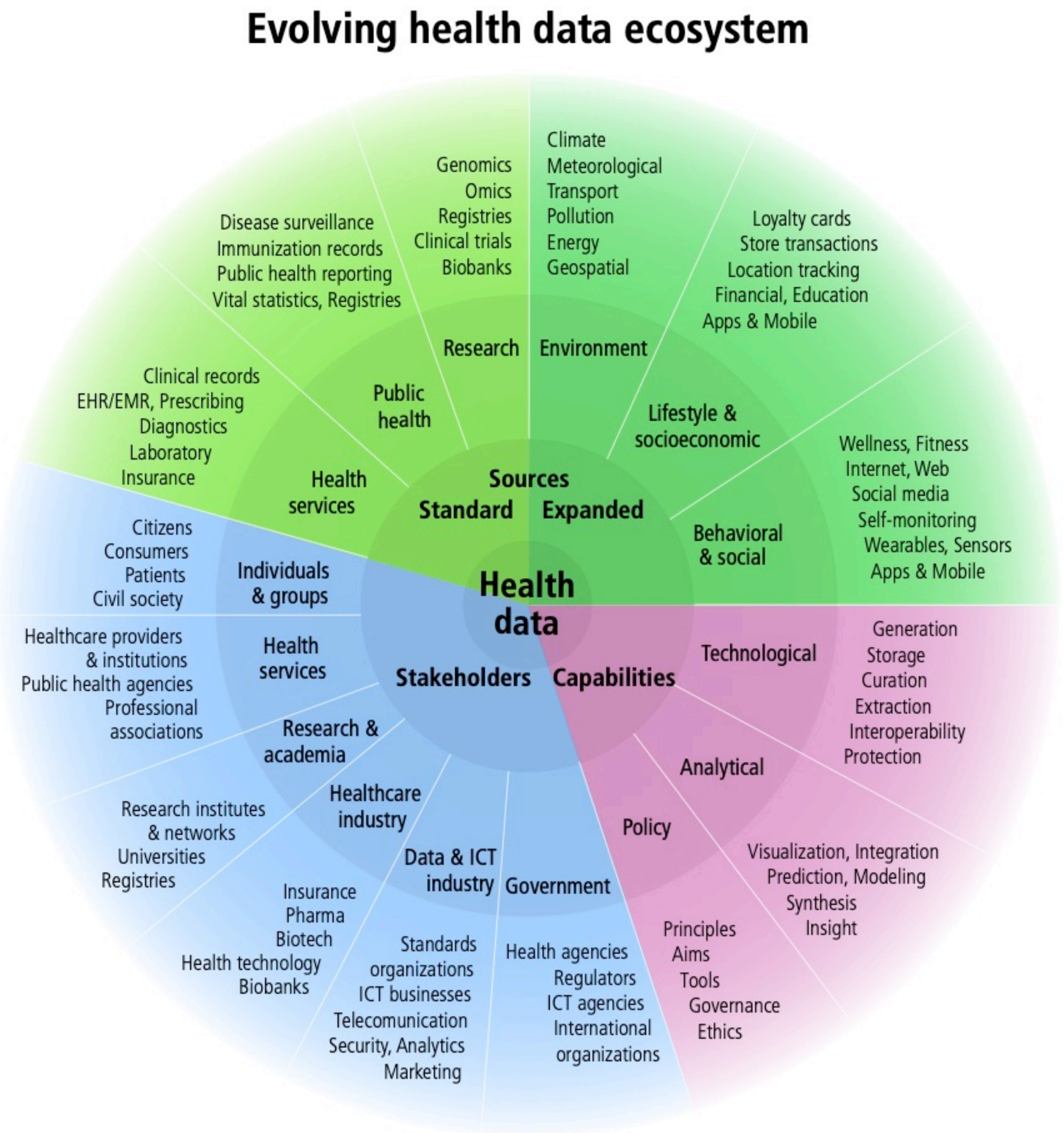


http://sites.ieee.org/futuredirections/2018/05/18/the-future-of-health-care-is-tied-to-big-data/

Big Data em Saúde



DADOS DE SAÚDE

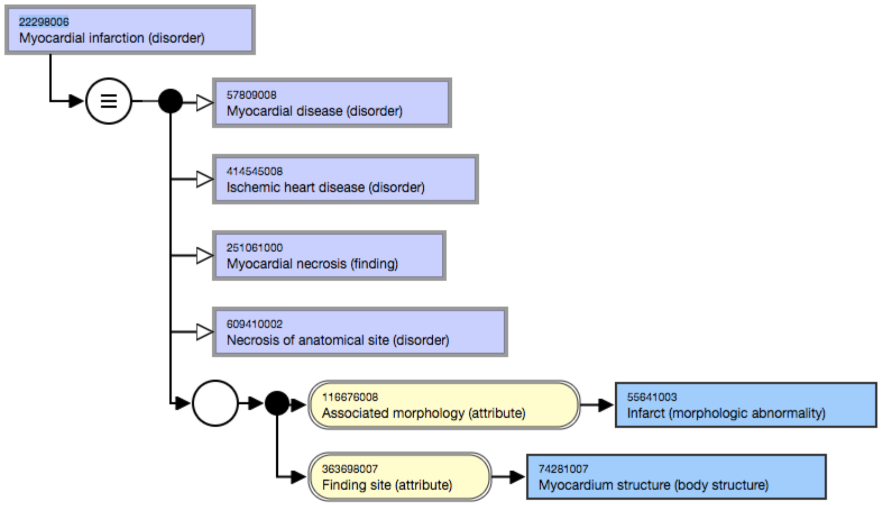


E. Vayena, J. Dzenowagis, M. Langfeld, 2016

Dados de saúde

1738.00 663v100,
J45, J44, 66YL.11,
G20.00, 662O.00,
1738.00 1682.00,
I50, 06,
116676008, I21.00

Structured



Orientation	plus
Stabilized	plus
Make rs174546(C;C)	
Make rs174546(C;T)	
Make rs174546(T;T)	
Reference	GRCh38 38.1/141
Chromosome	11
Position	61802358
Gene	FADS1
is a	snp
is	mentioned by
dbSNP	rs174546

0 Heart failure

Excl.: complicating:
- abortion or ectopic or molar pregnancy (O00-O07, O08.8)
- obstetric surgery and procedures (O75.4)

due to hypertension (I11.0)
- with renal disease (I13.-)

following cardiac surgery or due to presence of cardiac prosthesis (I97.1)
neonatal cardiac failure (P29.0)

0.0 Congestive heart failure
Congestive heart disease
Right ventricular failure (secondary to left heart failure)

0.1 Left ventricular failure
Cardiac asthma
Left heart failure
Oedema of lung | with mention of heart disease NOS or heart failure
Pulmonary oedema |

0.9 Heart failure, unspecified
Cardiac, heart or myocardial failure NOS

Dados de saúde

1738.00 663v100,
J45, J44, 66YL.11,
G20.00, 662O.00,
1738.00 1682.00,
I50, 06,
116676008, I21.00

Structured

220, 110, 0.002, 1,
200, 3, 2, 2.1, 2.01,
20, 1, 99092, 1.2,
99, 123, 6, 23.2,
878, 9901, 11,
203.1

Semi-structured

Dados de saúde

1738.00 663v100,
J45, J44, 66YL.11,
G20.00, 662O.00,
1738.00 1682.00,
I50, 06,
116676008, I21.00

Structured

220, 110, 0.002, 1,
200, 3, 2, 2.1, 2.01,
20, 1, 99092, 1.2,
99, 123, 6, 23.2,
878, 9901, 11,
203.1

Semi-structured

~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~

Unstructured

76 yo man with h/o HTN, DM, and sleep apnea who presented to the ED complaining of chest pain. He states that the pain began the day before and consisted of a sharp pain that lasted around 30 seconds, followed by a dull pain that would last around 2 minutes. The pain was located over his left chest area somewhat near his shoulder. The onset of pain came while the patient was walking in his home. He did not sit and rest during the pain, but continued to do household chores. Later on in the afternoon he went to the gym where he walked 1 mile on the treadmill, rode the bike for 5 minutes, and swam in the pool. After returning from the gym he did some work out in the yard, cutting back some vines. He did not have any reoccurrences of chest pain while at the gym or later in the evening. The following morning (of his presentation to the ED) he noticed the pain as he was getting out of bed. Once again it was a dull pain, preceded by a short interval of a sharp pain. The patient did experience some tingling in his right arm after the pain ceased. He continued to have several episodes of the pain throughout the morning, so his daughter-in-law decided to take him to the ED around 12:30pm. The painful episodes did not increase in intensity or severity during this time.

Dados de saúde

1738.00 663v100, J45, J44, 66YL.11, G20.00, 662O.00, 1738.00 1682.00, I50, 06, 116676008, I21.00

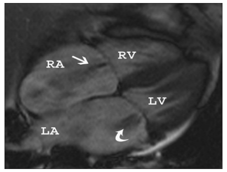
Structured

220, 110, 0.002, 1, 200, 3, 2, 2.1, 2.01, 20, 1, 99092, 1.2, 99, 123, 6, 23.2, 878, 9901, 11, 203.1

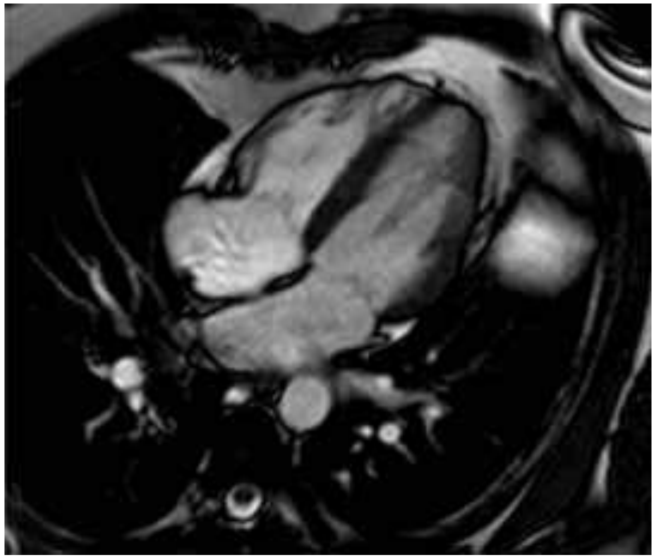
Semi-structured

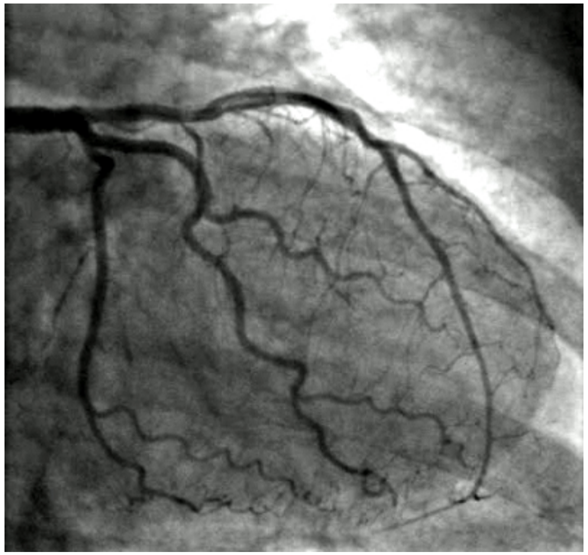
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~

Unstructured



Binary





Dados de saúde

1738.00 663v100, J45, J44, 66YL.11, G20.00, 662O.00, 1738.00 1682.00, I50, 06, 116676008, I21.00

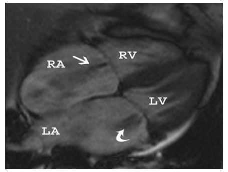
Structured

220, 110, 0.002, 1, 200, 3, 2, 2.1, 2.01, 20, 1, 99092, 1.2, 99, 123, 6, 23.2, 878, 9901, 11, 203.1

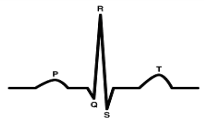
Semi-structured

~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~
~~~~~~~~~~~~

Unstructured



Binary



<u>Streaming</u>

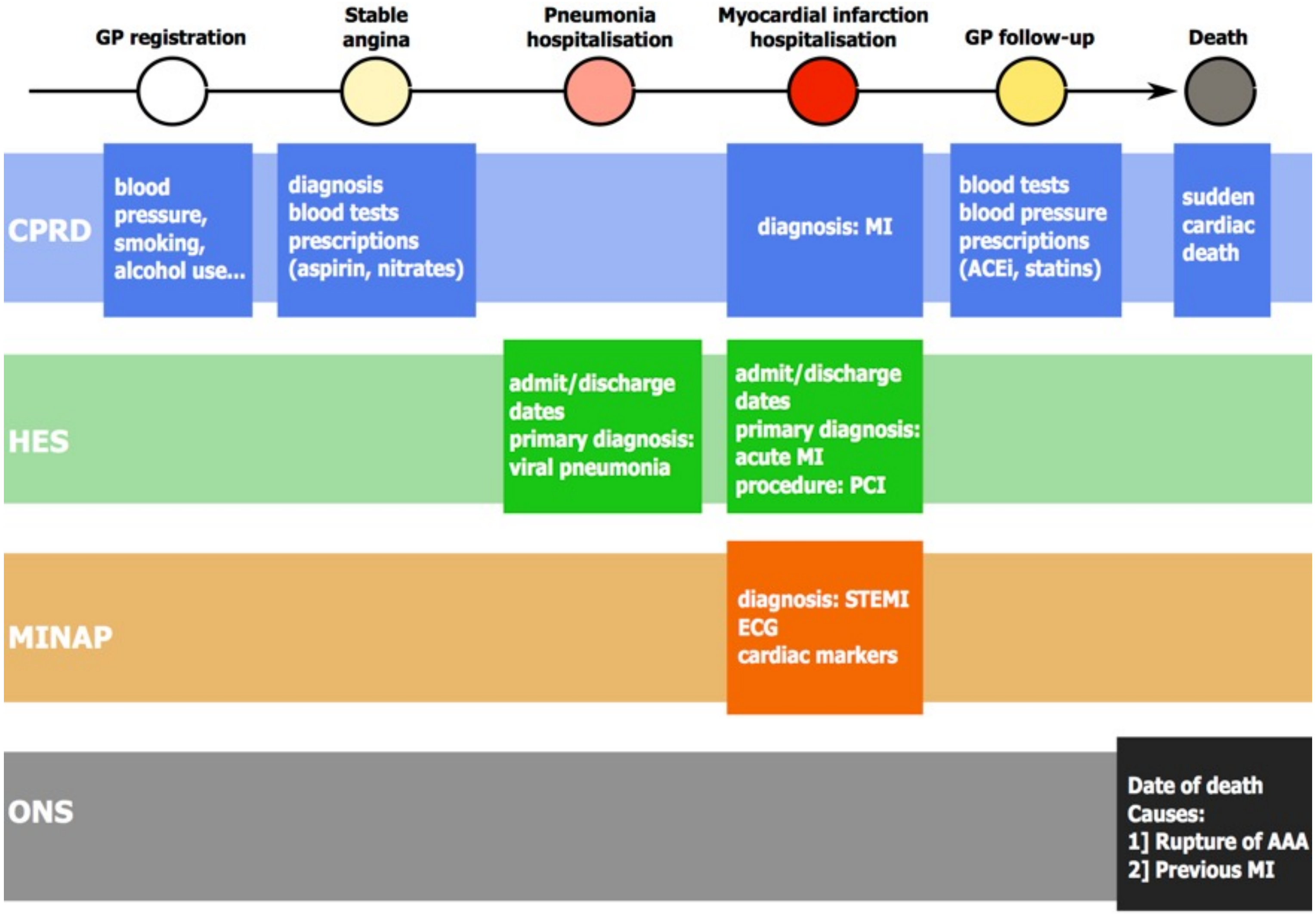
APLICAÇÕES

- ✓ Estudos epidemiológicos de base populacional (coorte)
- ✓ Análises clínicas
- ✓ Descoberta de novos fármacos
- ✓ Cura de doenças / resistência antimicrobiana
- ✓ Medicina personalizada
- ✓ Prevenção de visitas desnecessárias ao médico

- ✓ Análise preditiva em saúde
- ✓ Registros eletrônicos (*electronic health records – EHR*)
- ✓ Monitoramento em tempo real
- ✓ Sistemas de suporte à decisão
- ✓ Medicina assistida por computação / *mobile health* (m-Health)

Aplicações

Population-based clinical epidemiology



Denaxas S. *et al.*, Int J Epidemiology, 2013, doi: 10.1093/ije/dys188

Aplicações

Type-2 Diabetes and 12 CVDs

Initial presentation of cardiovascular disease	Number of events: No diabetes	Number of events: Type 2 diabetes	Hazard ratio (95% CI)	p value
Stable angina	12 232	728	1·62 (1·49–1·77)	<0·0001
Unstable angina	5286	245	1·53 (1·32–1·76)	<0·0001
Non-fatal myocardial infarction	15 191	706	1·54 (1·42–1·67)	<0·0001
Unheralded coronary death	5101	255	1·43 (1·23–1·65)	<0·0001
Heart failure	13 072	866	1·56 (1·45–1·69)	<0·0001
Arrhythmia or sudden cardiac death	3218	100	0·95 (0·76–1·19)	0·65
Transient ischaemic attack	10 990	513	1·45 (1·31–1·60)	<0·0001
Ischaemic stroke	5643	316	1·72 (1·52–1·95)	<0·0001
Subarachnoid haemorrhage	1260	11	0·48 (0·26–0·89)	0·020
Intracerebral haemorrhage	2265	84	1·28 (1·02–1·62)	0·035
Peripheral arterial disease	10 074	992	2·98 (2·76–3·22)	<0·0001
Abdominal aortic aneurysm	3051	62	0·46 (0·35–0·59)	<0·0001

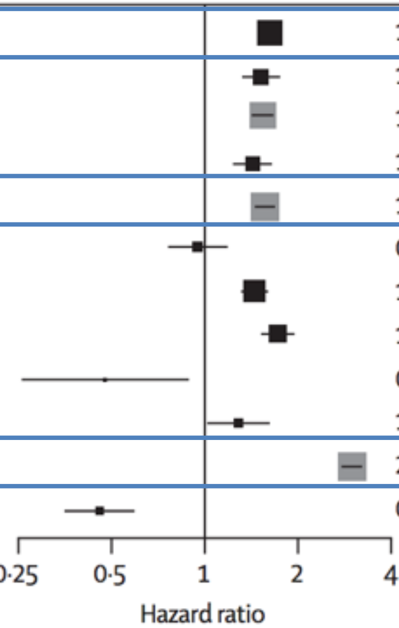


Figure 3: Association of type 2 diabetes with 12 cardiovascular diseases in patients aged ≥30 years

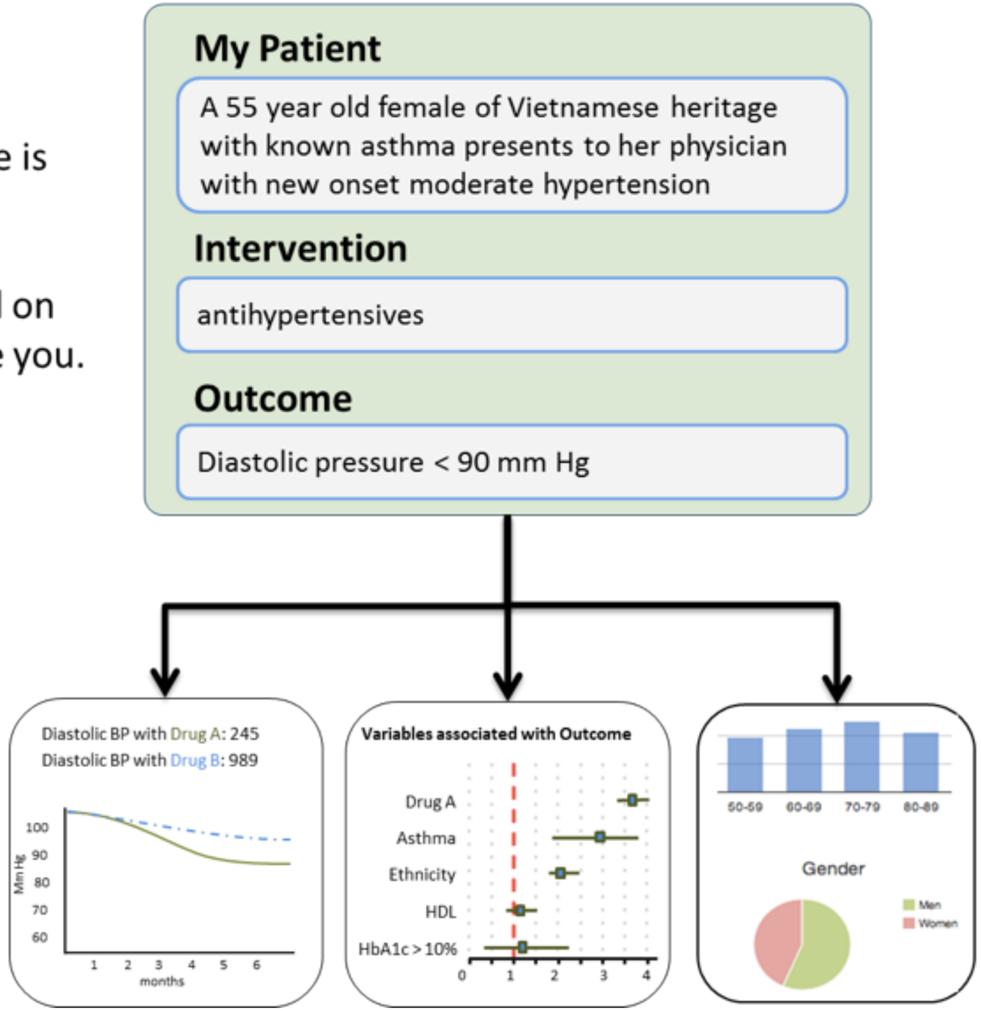
Shah A. *et al.*, Lancet Diab Endocrinol, 2015, doi: 10.1016/S2213-8587(14)70219-0

Aplicações

Clinical trials pipelines

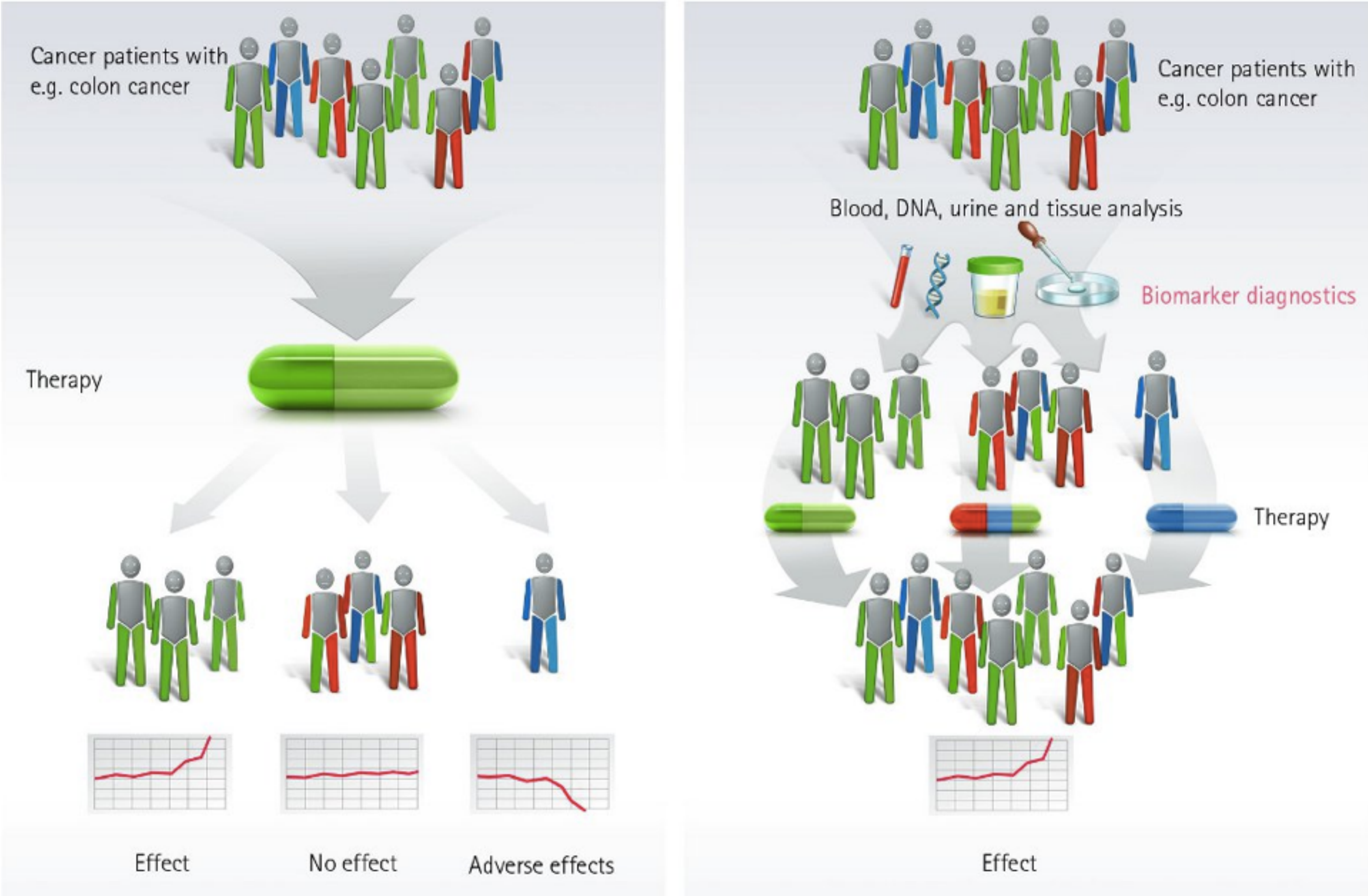
Problem: A lot of medical care is educated guesses

Opportunity: Decisions based on what happened to people like you.



Longhurst A. A. *et al.*, Health Affairs, 2014, doi:10.1377/hlthaff.2014.0099

Aplicações

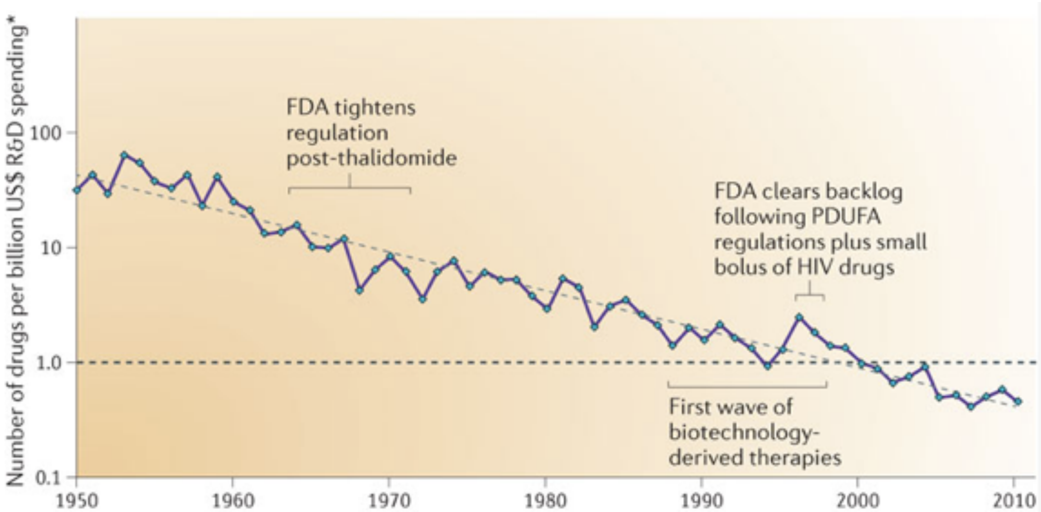


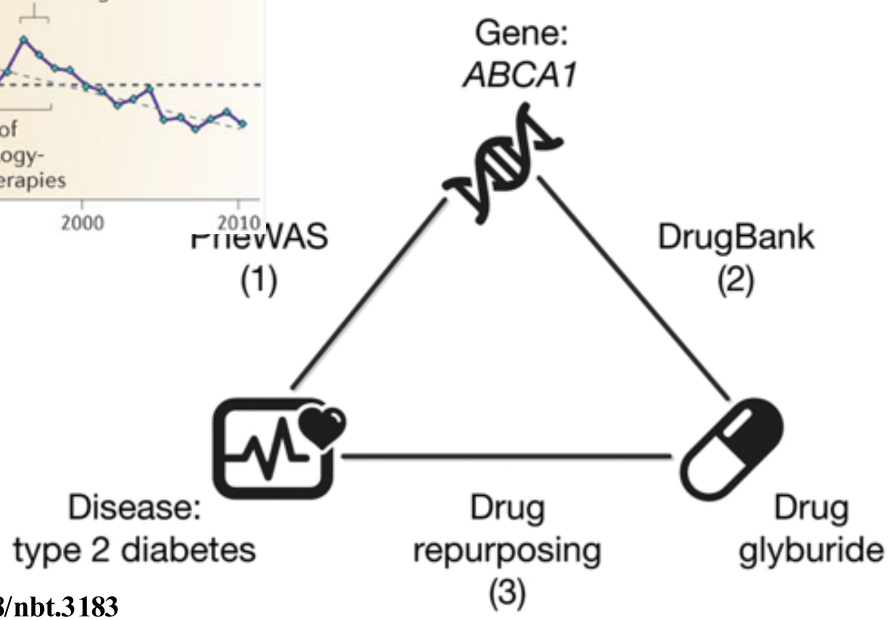
Precision medicine

Aplicações

Drug discovery and repositioning

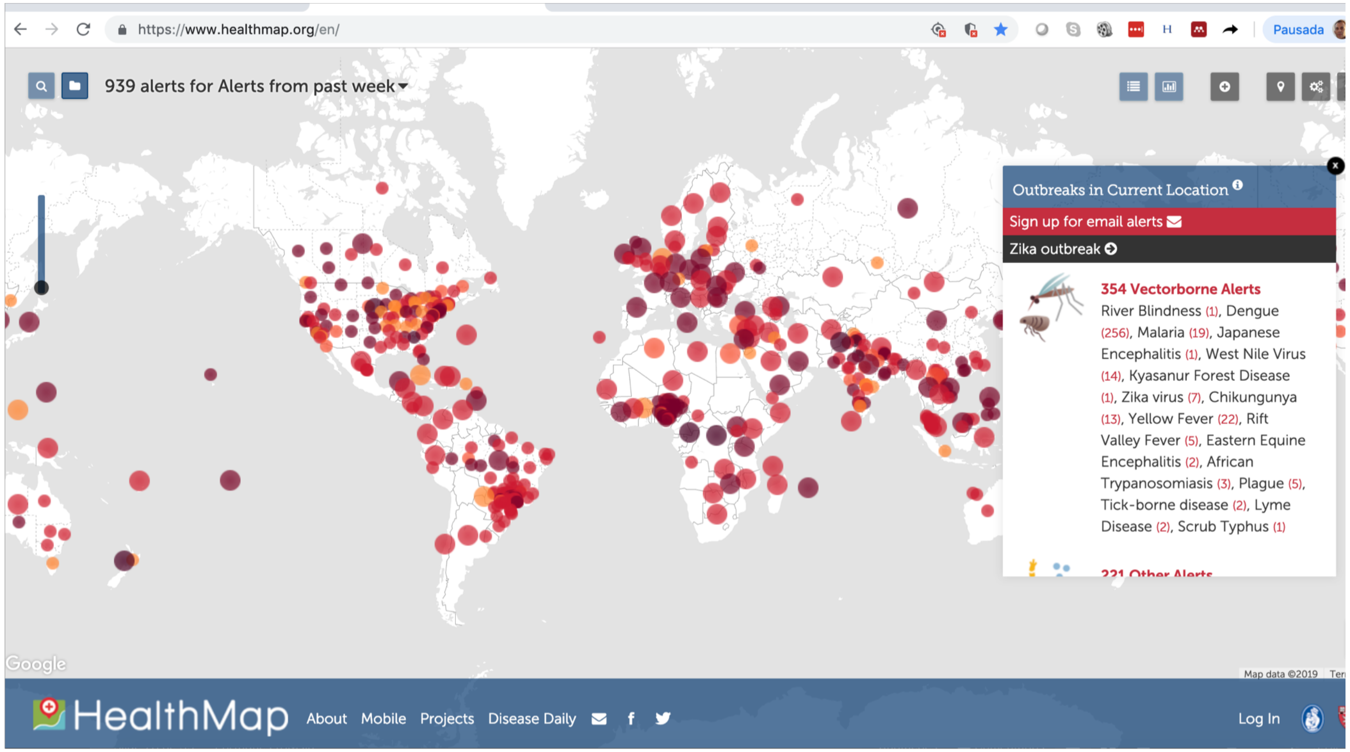
Challenge: costs (5-11bn USD), time (17 years)

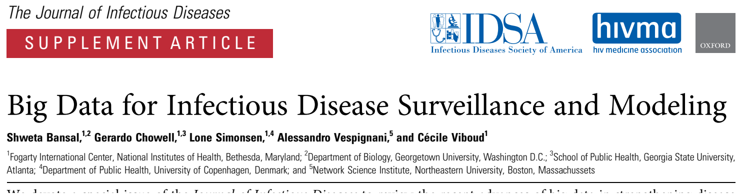




Rastergar-Mojarad M. *et al.*, Nature Biotech, 2015, doi: 10.1038/nbt.3183
Scannell J.W. *et al.*, Nat Review Drug Discovery, 2012, doi: 10.1038/nrd3681

Aplicações




The Journal of Infectious Diseases

SUPPLEMENT ARTICLE

Big Data for Infectious Disease Surveillance and Modeling

Shweta Bansal,[1,2] Gerardo Chowell,[1,3] Lone Simonsen,[1,4] Alessandro Vespignani,[5] and Cécile Viboud[1]

[1]Fogarty International Center, National Institutes of Health, Bethesda, Maryland; [2]Department of Biology, Georgetown University, Washington D.C.; [3]School of Public Health, Georgia State University, Atlanta; [4]Department of Public Health, University of Copenhagen, Denmark; and [5]Network Science Institute, Northeastern University, Boston, Massachusetts

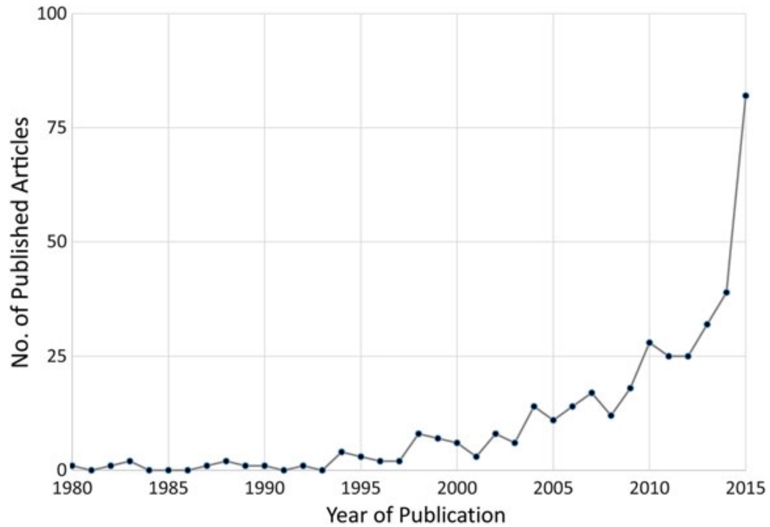
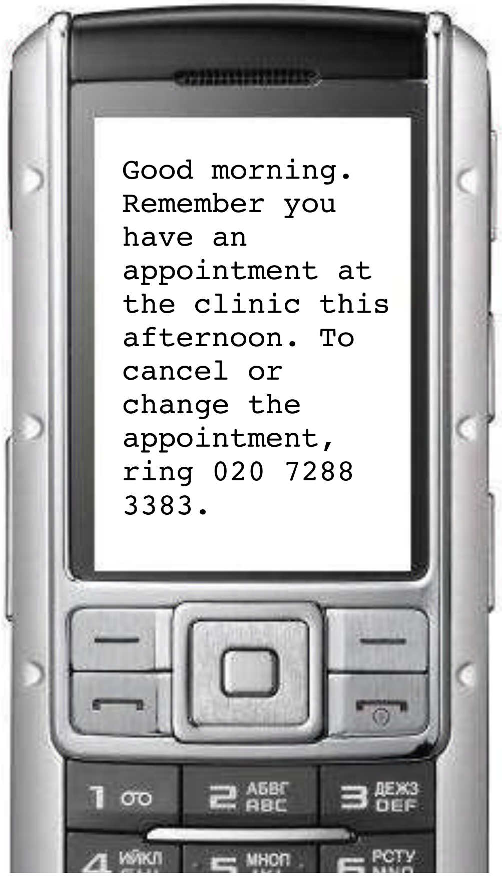


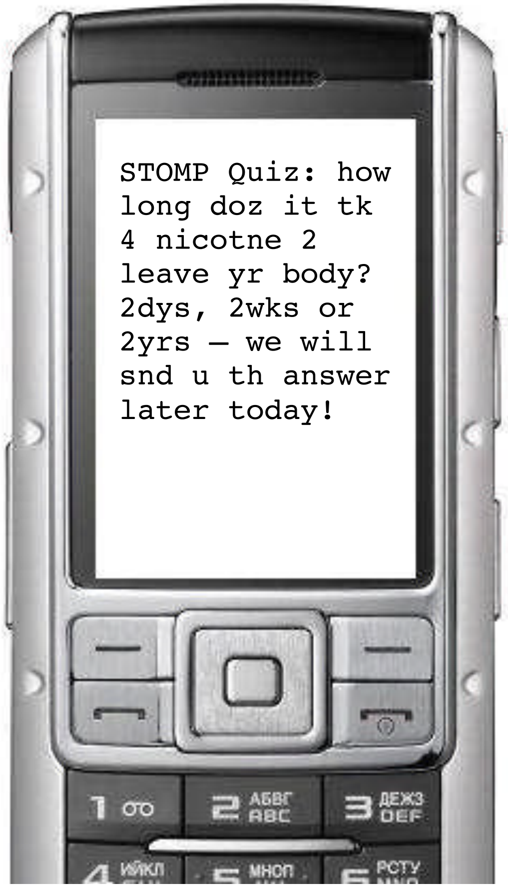
Figure 1. Exponential increase since the early 2000s in publications at the intersection of big data and infectious diseases. Annual trends in the number of publications were identified through a Scopus search for articles published between 1980 and 2015, using the following keywords: (big data AND infectious diseases) OR (big data AND epidemics) OR (digital epidemiology AND infectious diseases).

Aplicações m-Health

Dr Henry Potts: h.potts@ucl.ac.uk



Appointment reminder

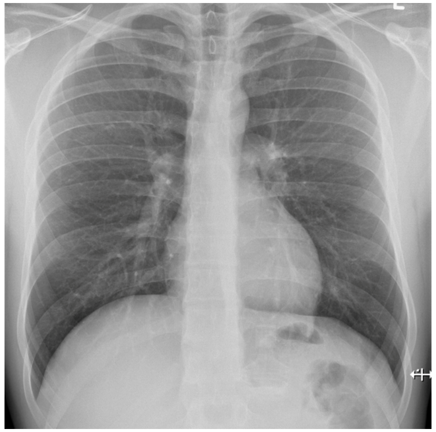


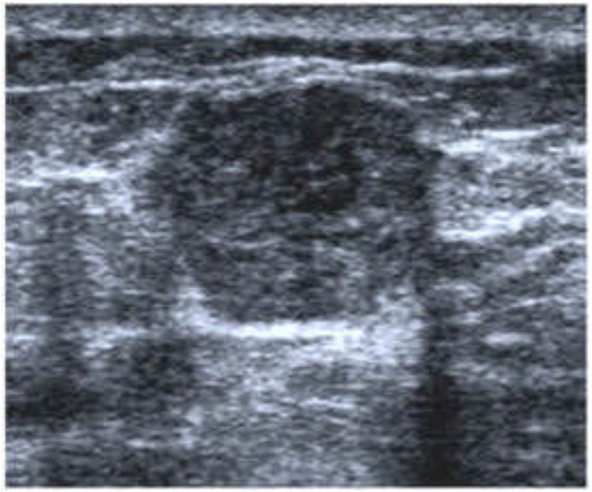
Behaviour change

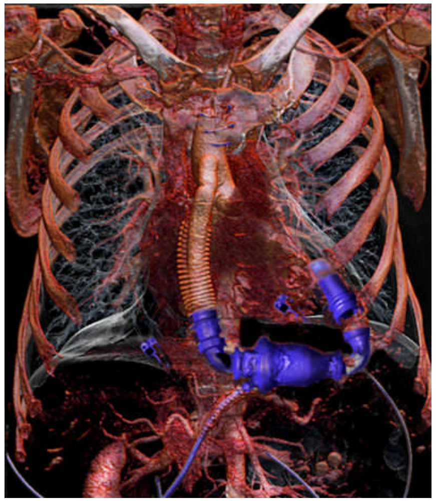
Personal health record

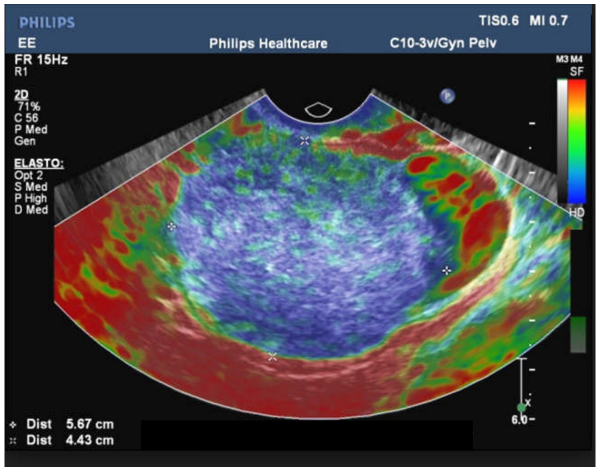
Aplicações

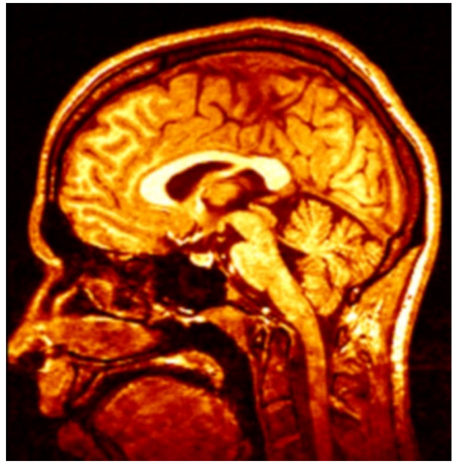
Medical image processing





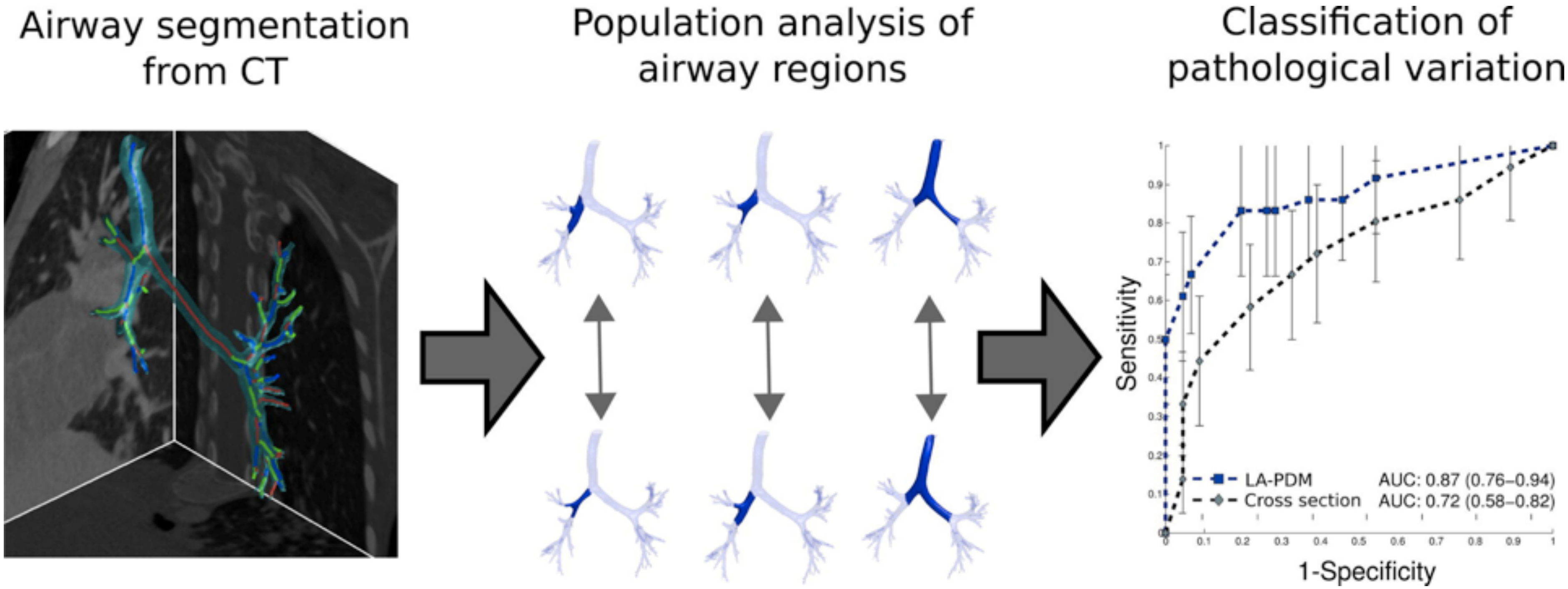






Aplicações

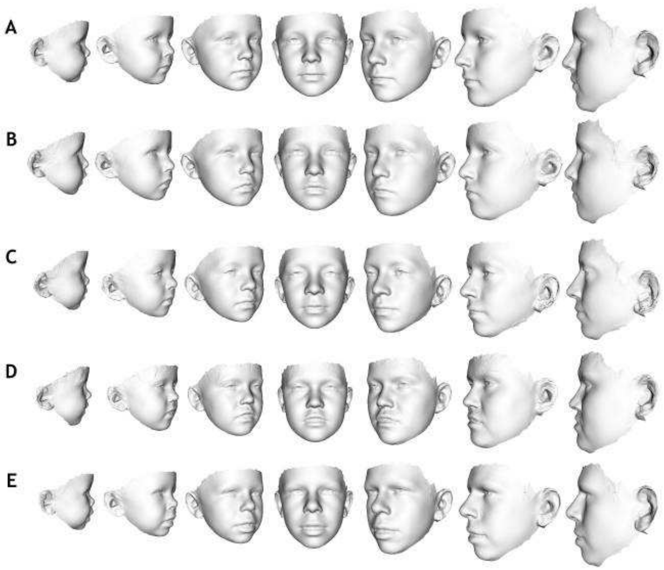
Medical image processing



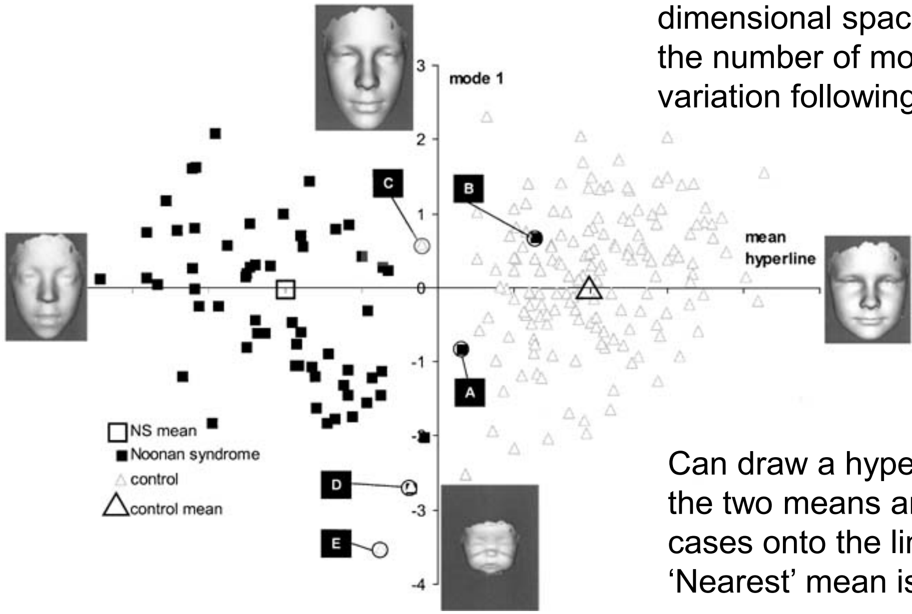
Aplicações

Hammond et al. Using a database of 3D facial scans to identify genetic disorders

Each row shows a series of views of the mean face for a different syndrome



Each face is a point in n dimensional space where n is the number of modes of variation following PCA



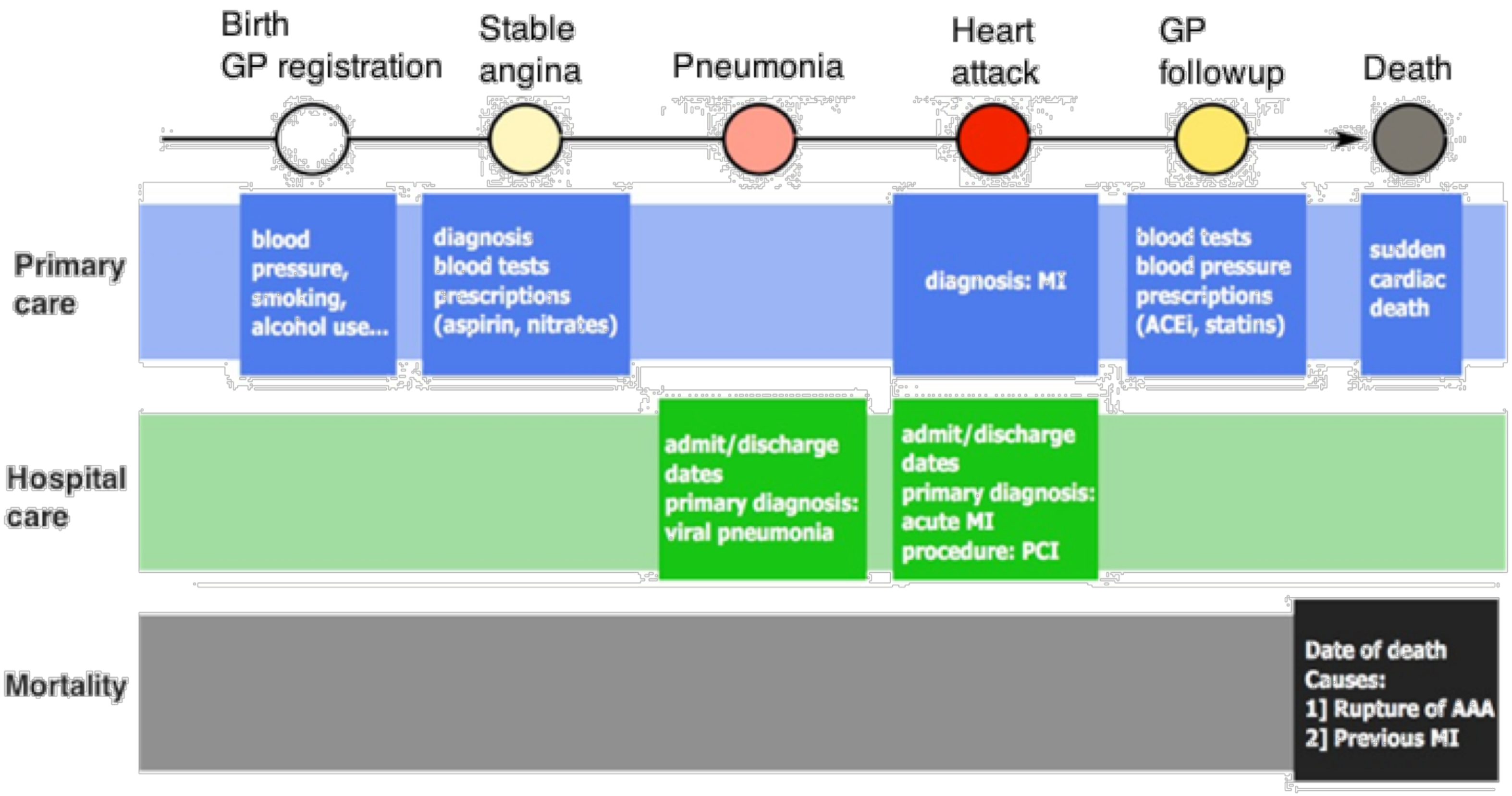
Can draw a hyperline between the two means and project cases onto the line
'Nearest' mean is one classifier

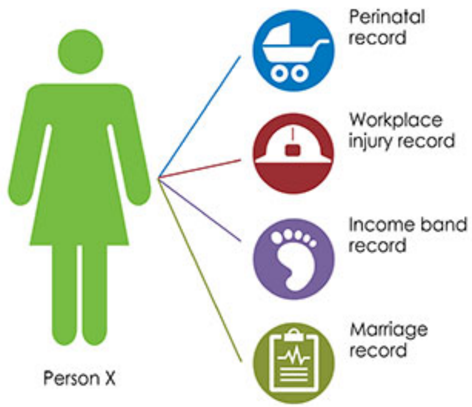
DESAFIOS

CASO CLÍNICO
Caminho do Paciente
MINISTÉRIO DA SAÚDE
PROADI SUS
OSWALDO CRUZ
HOSPITAL ALEMÃO
IHE
Portal de Resultados
VISUALIZADOR
ID CLÍNICO
1
Queixa de diminuição da visão
Encaminhado ao oftmologista.
Resultados de Exames
3
Diagnóstico de micro-aneurisma e retinopatia diabética.
Lab. AE RAC
5
Internação de Emergência.
AB LAB AE RAC
Autorizou Consentimento
AB (Unidade de Atenção Básica)
AB
CONSENTIMENTO
Laboratório (realização de exames)
AE (Atendimento Especializado de Oftamologista)
24H
CONSENTIMENTO
AB (Unidade de Atenção Básica)
AB
Pronto Socorro (Cirurgia de Emergência)
24H
24H
AE (Atendimento Especializado)
Visualizar Registros
RAC
2
Realização de Exames.
Lab. AB RAC
4
Análise de resultados e medicação.
AB LAB AE S. A
6
Pós-Operatório com Cardiologista.
Parceiros Tecnológicos:
IHE
Legenda:
CONSENTIMENTO
Autorização de Consentimento
24H
Consentimento de 24 horas
Recuperar Informação
Enviar Informação

Vinculação de dados (*data linkage*)

- To bring together electronic records containing information from different sources about an entity (individual, organisation, location etc).



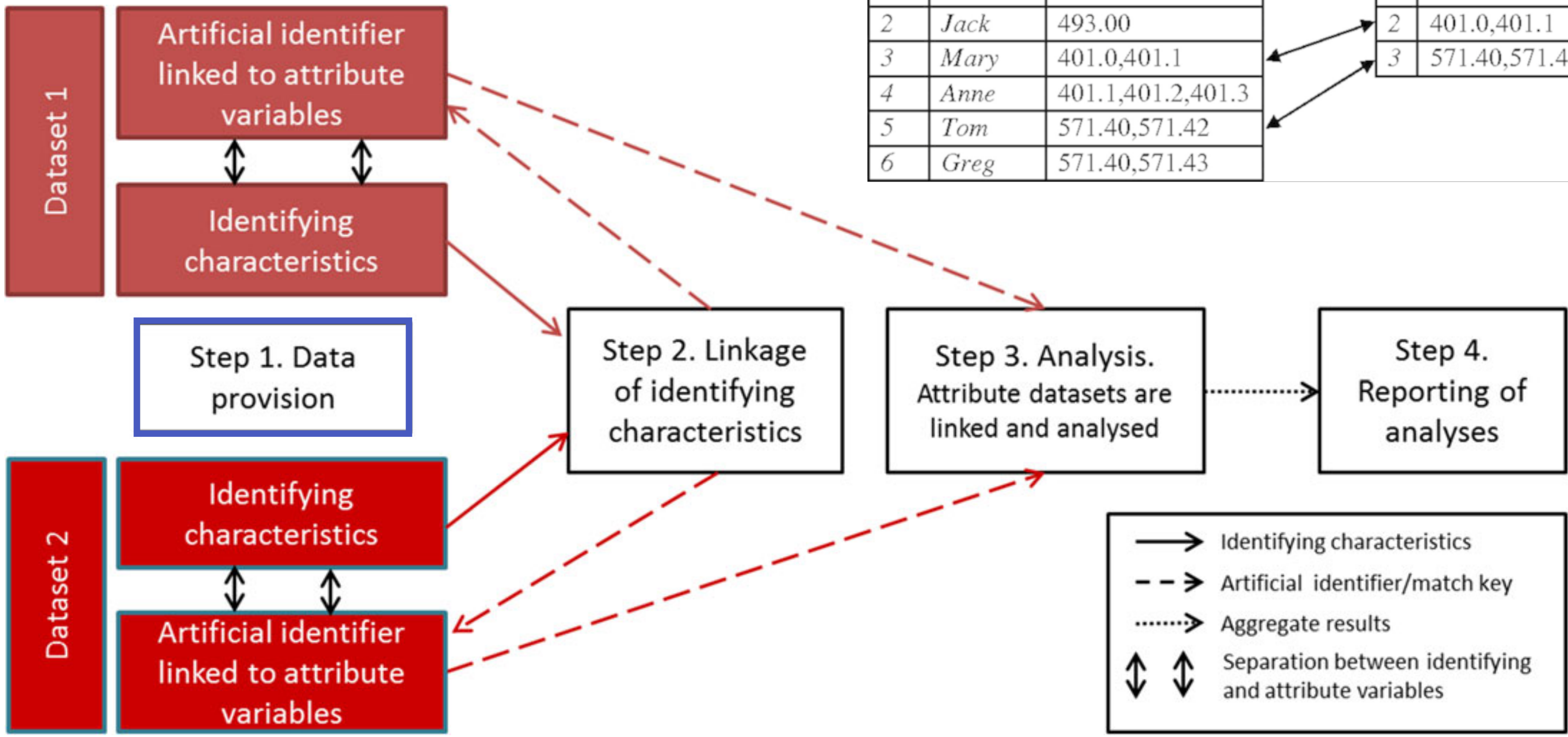


Vinculação de dados (*data linkage*)

- 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

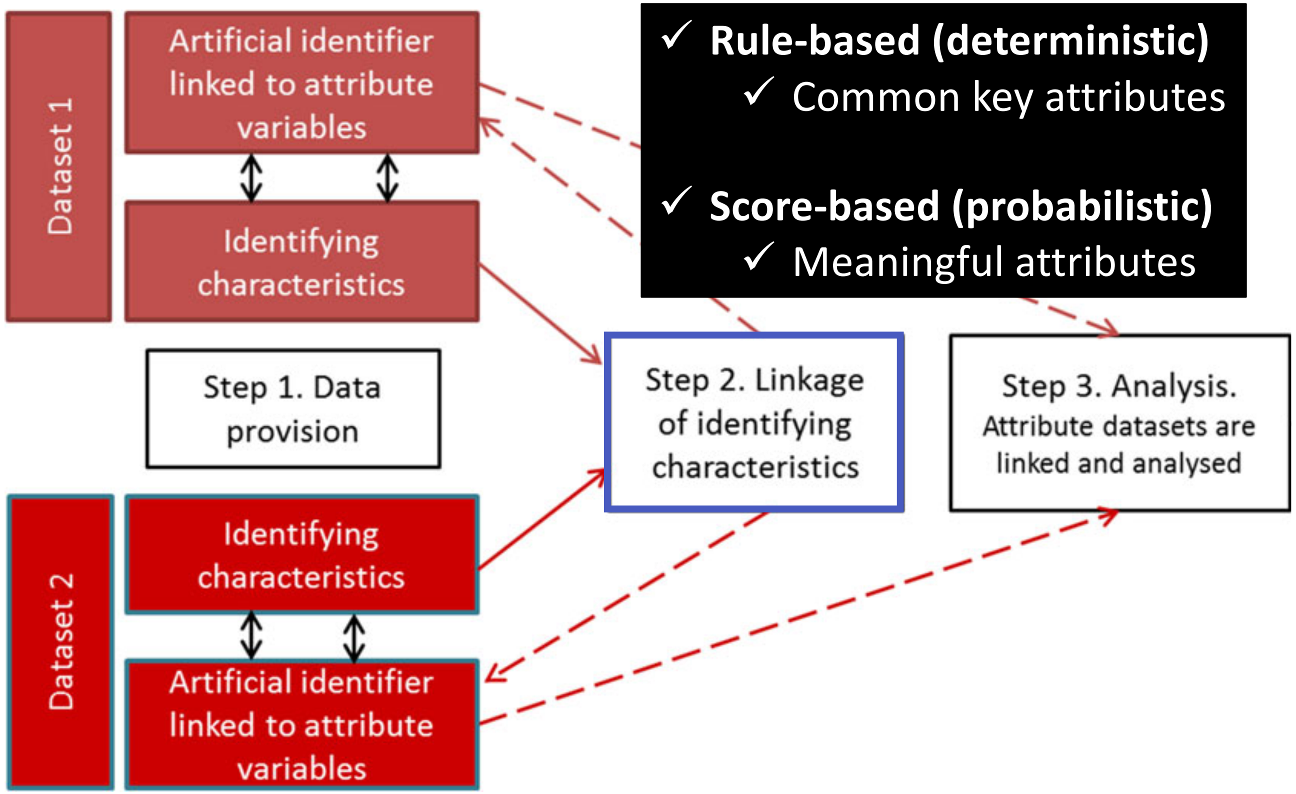
De-identified Research data (S)		
j	ICD9	DNA
1	493.00	*CT…A*
2	401.0,401.1	*AC…T*
3	571.40,571.42	*GC…A*

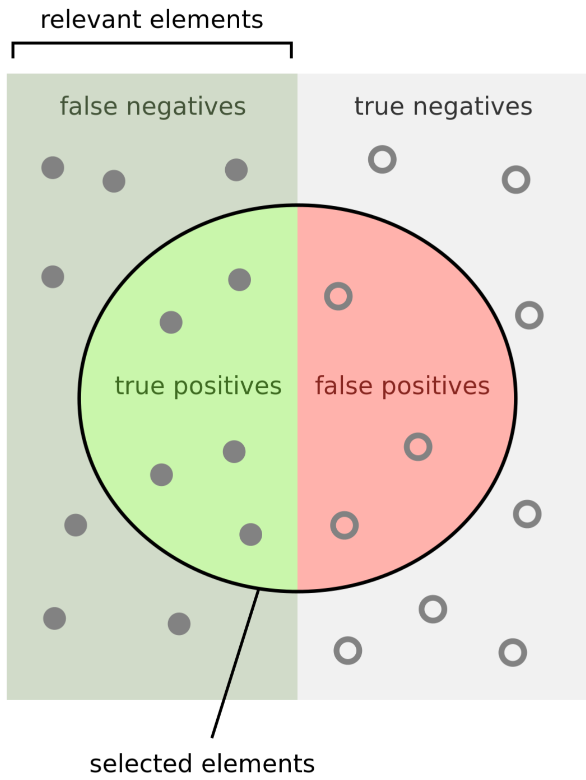


GUILD: GUidance for Information about Linking Data sets
Journal of Public Health , doi:10.1093/pubmed/fdx037

Vinculação de dados (*data linkage*)

- Pairwise comparison

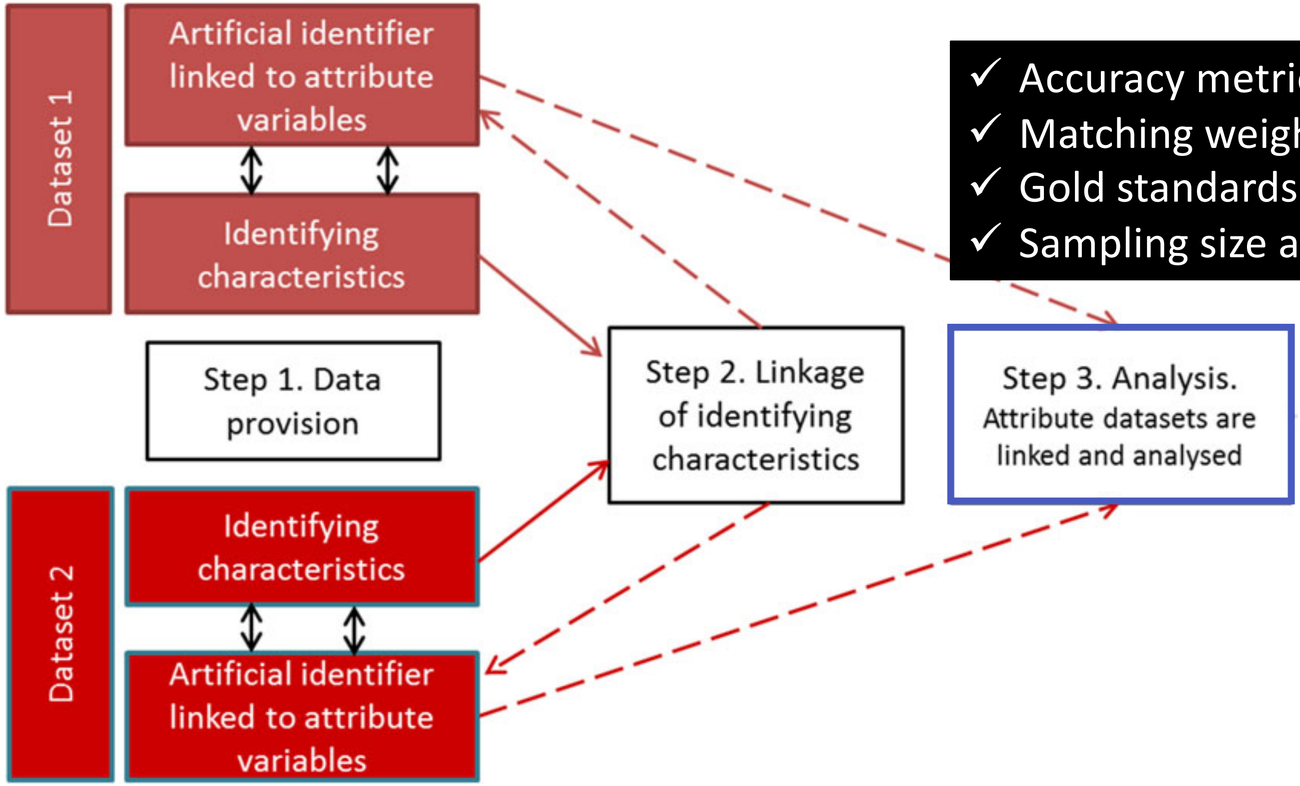




GUILD: GUidance for Information about Linking Data sets
Journal of Public Health , doi:10.1093/pubmed/fdx037

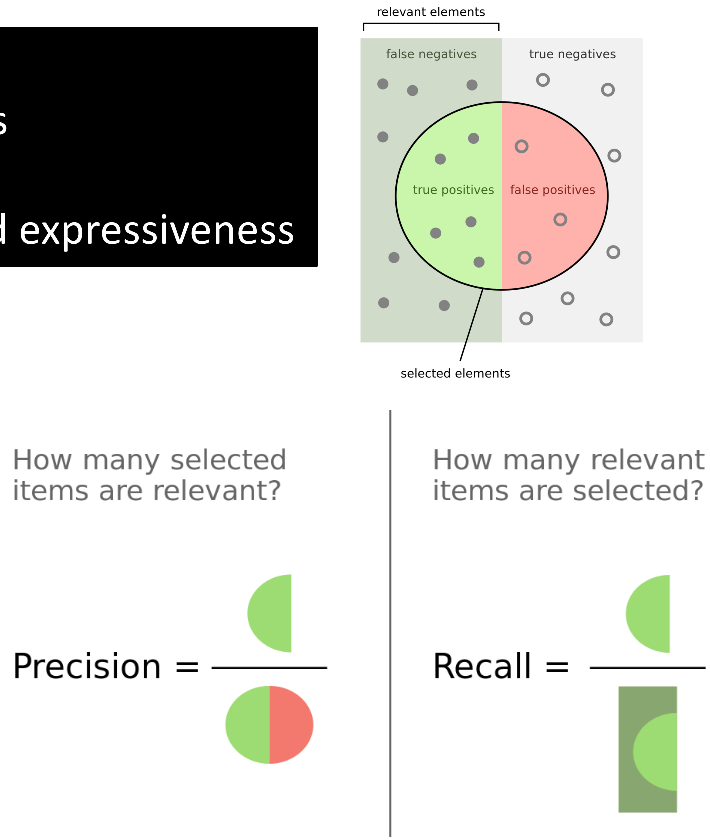
Vinculação de dados (*data linkage*)

- Accuracy ascertainment

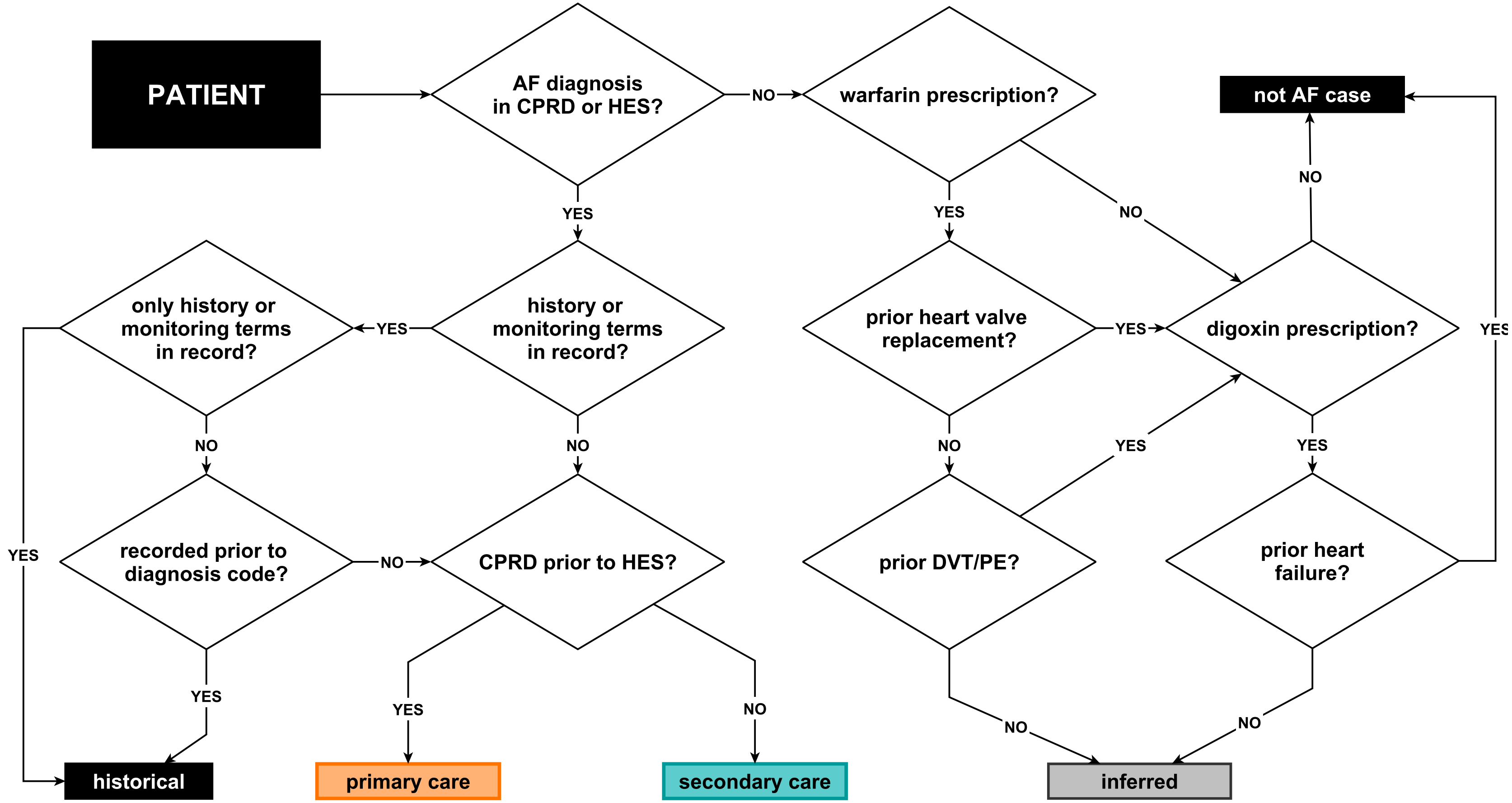


✓ Accuracy metrics
✓ Matching weights
✓ Gold standards
✓ Sampling size and expressiveness

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



Extratificação de pacientes (*phenotyping*)

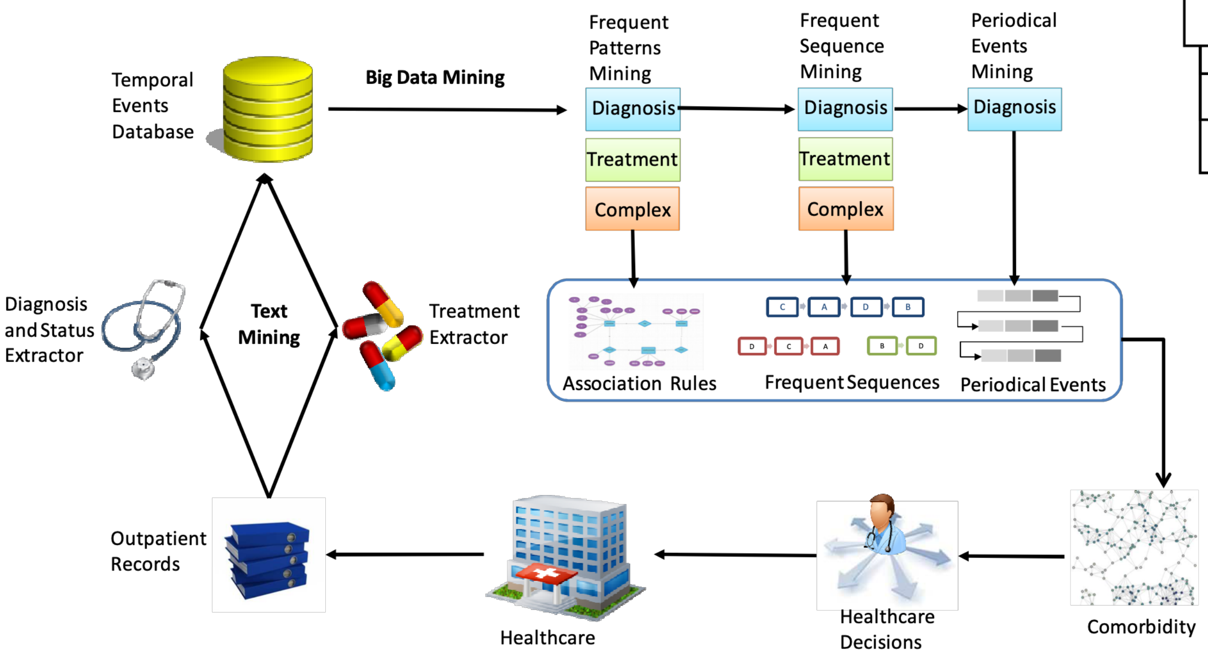


Morley K. *et al.*, PLOS ONE, 2014, doi: 10.1371/journal.pone.0110900

Análise de dados temporais

Big Data Analytics in Healthcare – Pattern Mining of Temporal Clinical Events

Svetla Boytcheva[1], Galia Angelova[1], Dimitar Tcharaktchiev[2], Zhivko Angelov[3]



healthcare

Review

A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining

Md Saiful Islam [1], Md Mahmudul Hasan [1], Xiaoyi Wang [1], Hayley D. Germack [1,2,3] and Md Noor-E-Alam [1,*]

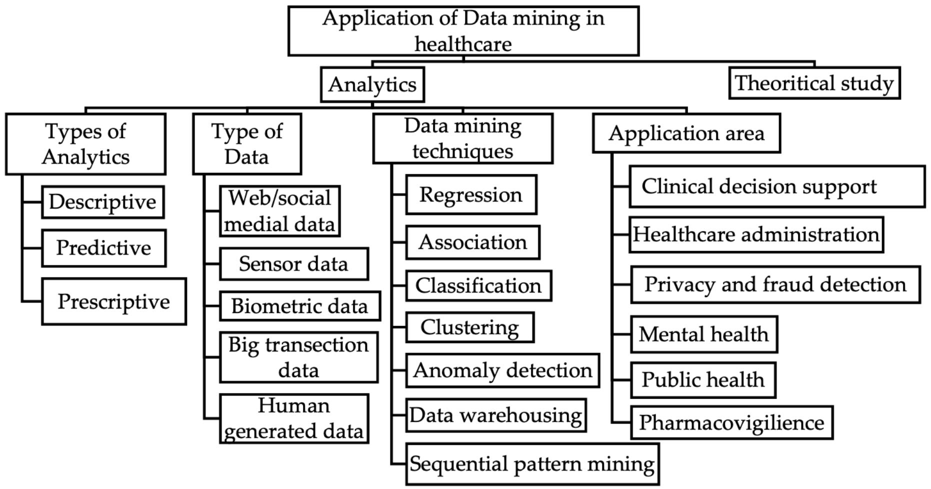


Figure 3. Classification scheme of the literature.

Análise preditiva

IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. 21, NO. 1, JANUARY 2017

Deep Learning for Health Informatics

Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang, *Fellow, IEEE*

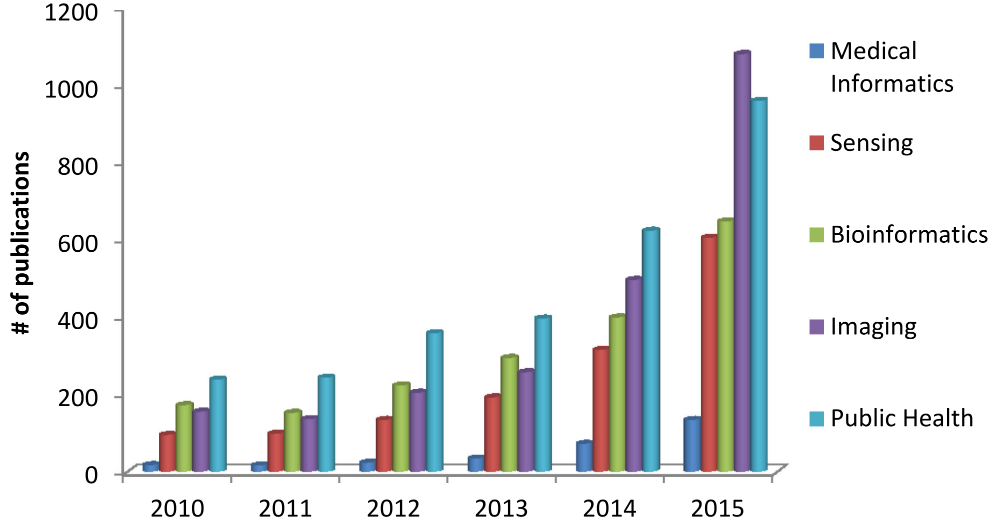
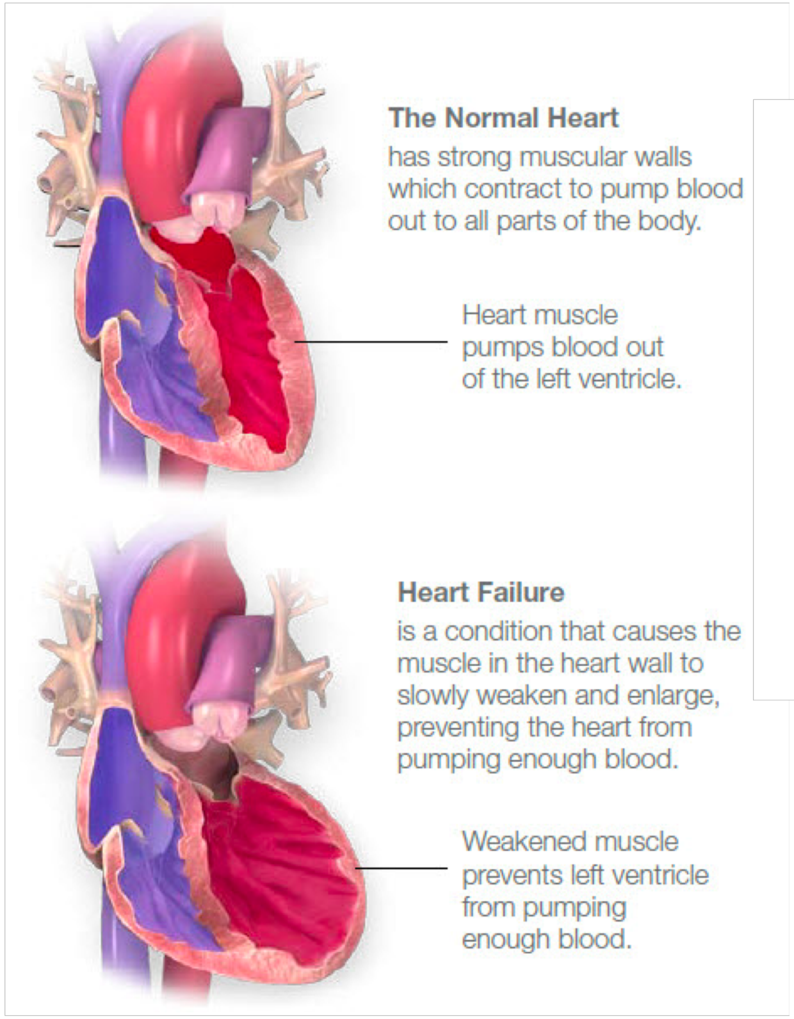
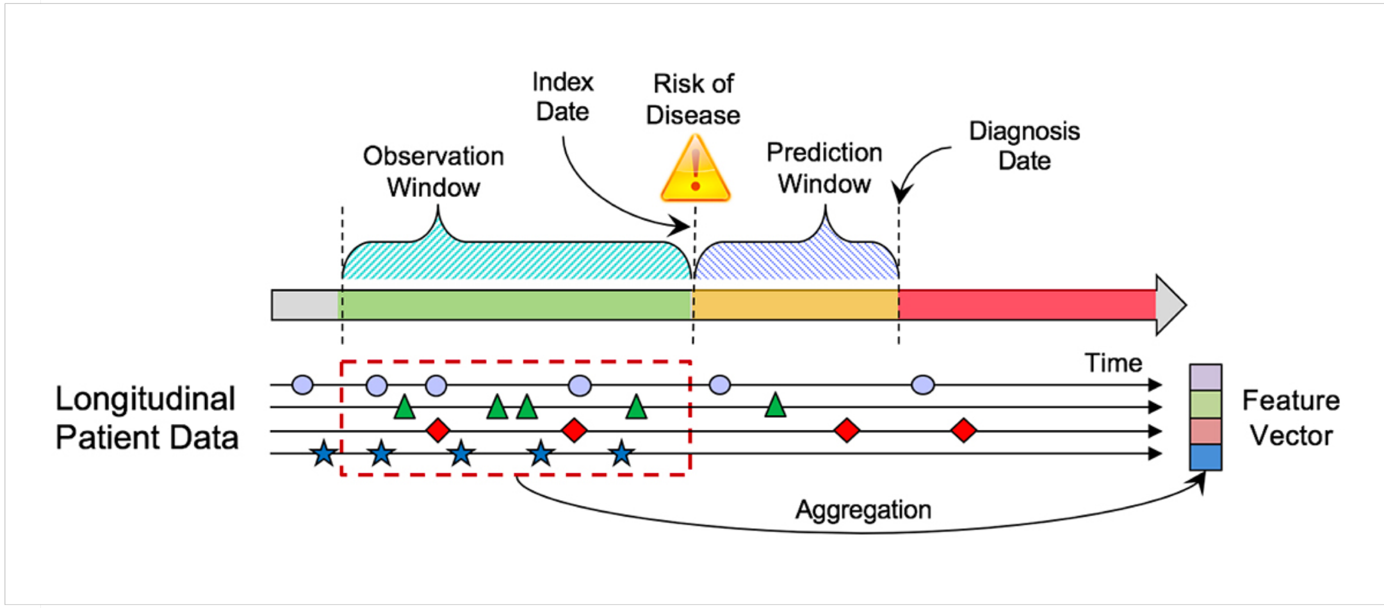


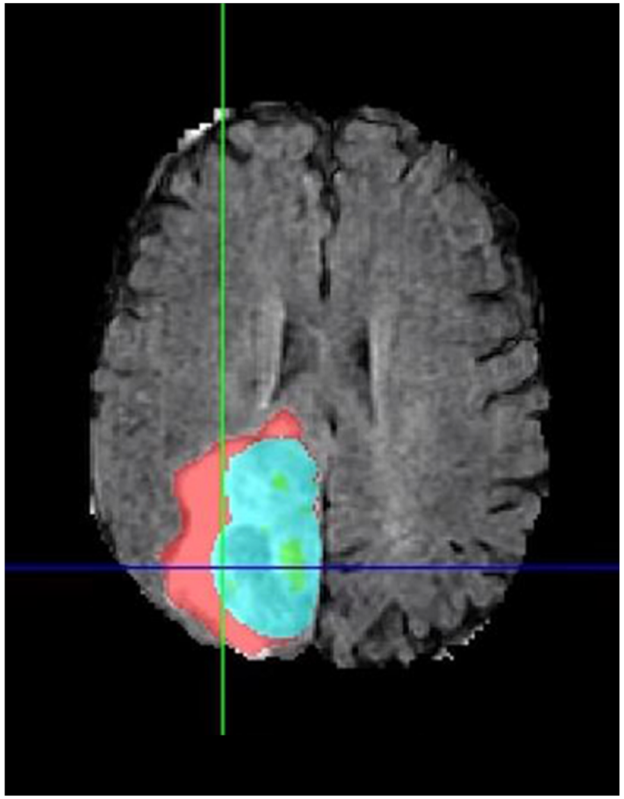
Fig. 1. Distribution of published papers that use deep learning in subareas of health informatics. Publication statistics are obtained from Google Scholar; the search phrase is defined as the subfield name with the exact phrase *deep learning* and at least one of *medical* or *health* appearing, e.g., "public health" "deep learning" medical OR health.

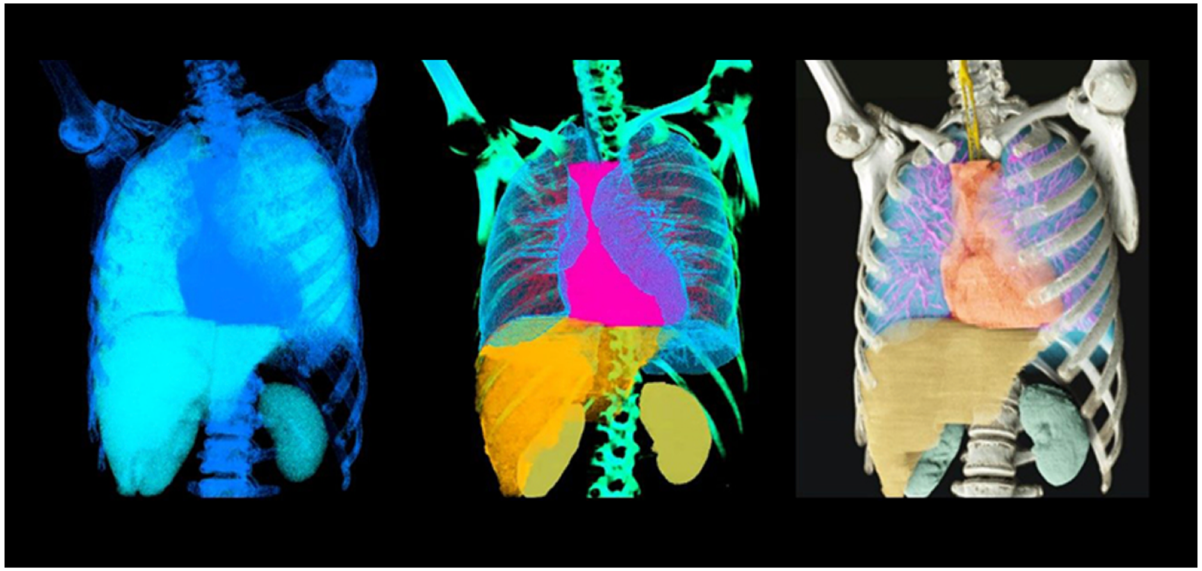
Análise preditiva





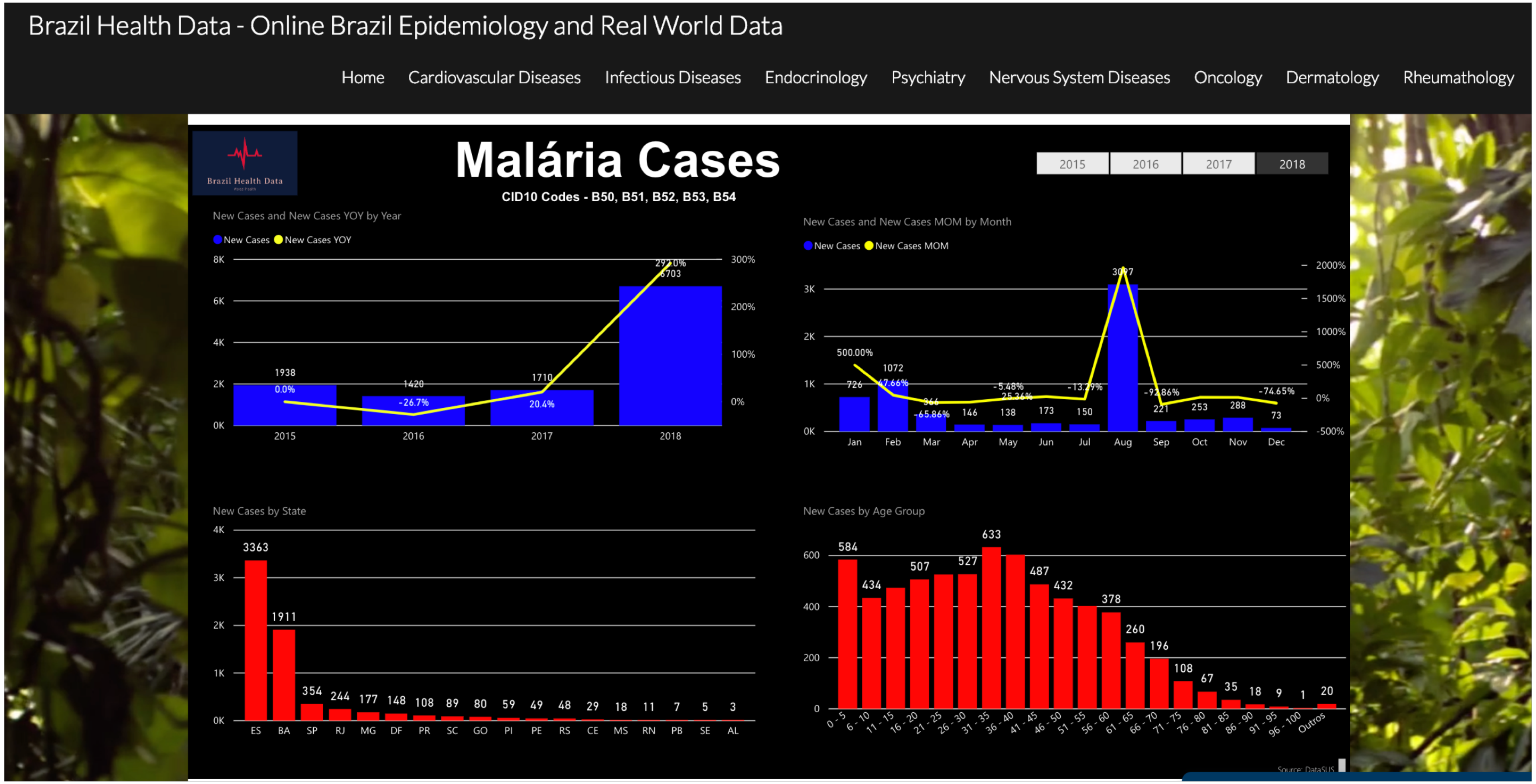
Processamento de imagens





Mineração visual de dados

https://www.brazilhealthdata.com/



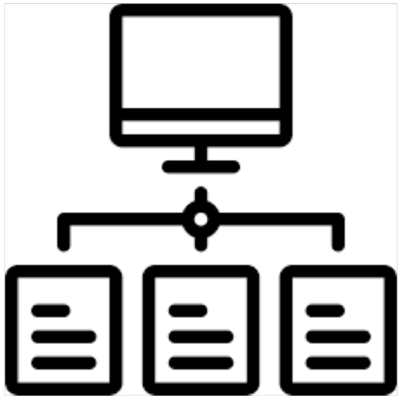
AtyImo

www.atyimolab.ufba.br

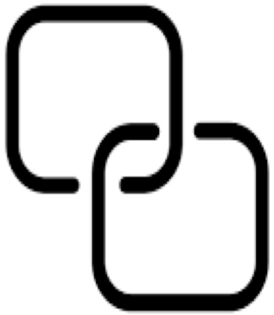
Atylmo

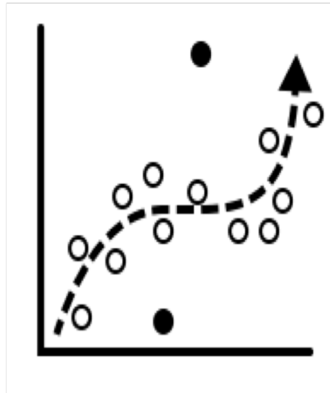
What we do

cloud
robotics



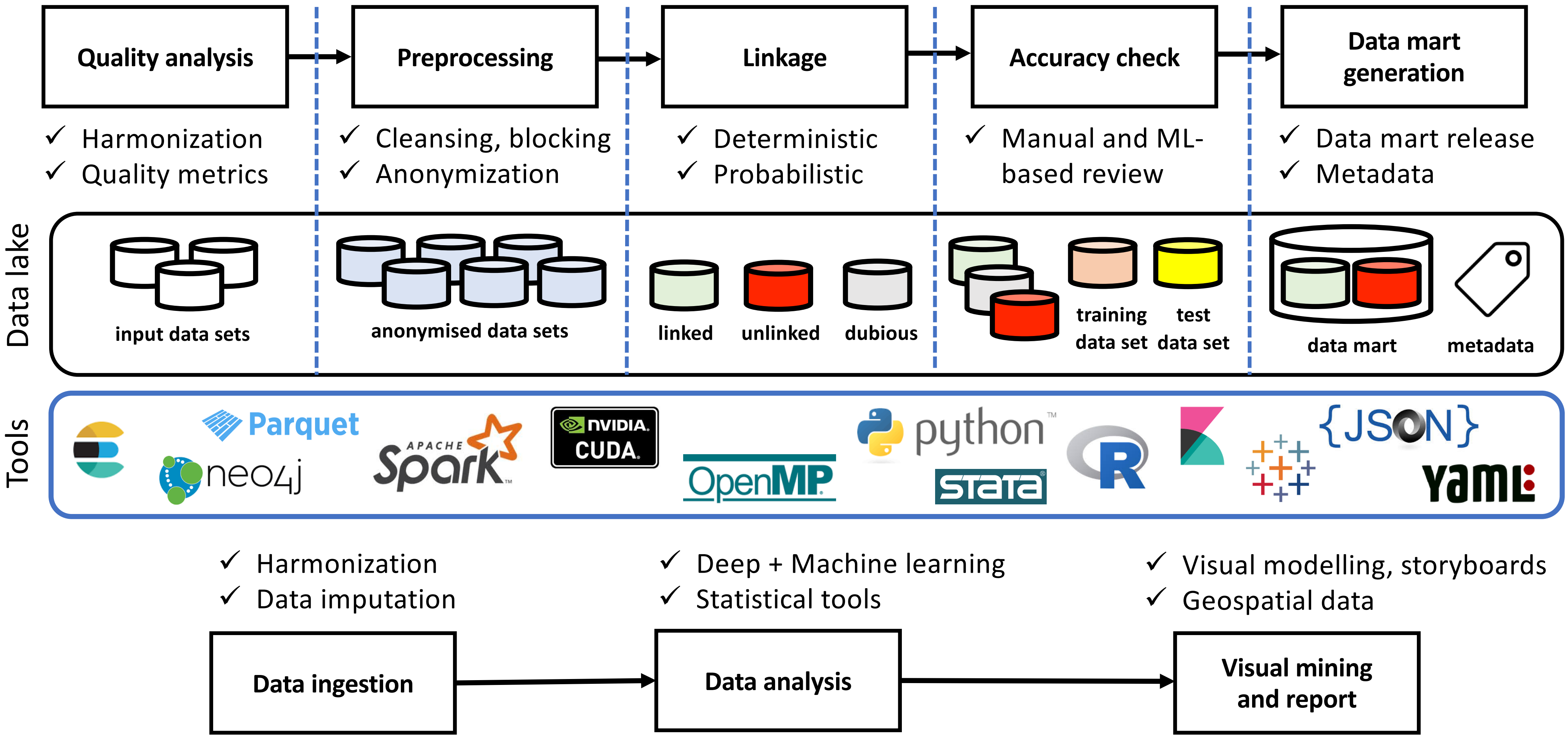
hybrid
parallel
computing





big data
linkage &
analytics

Atylmo – Data linkage platform



Data analytics pipeline

Brazilian governmental databases

Social programmes

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

≅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 we contribute to

✓ 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.

✓ 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

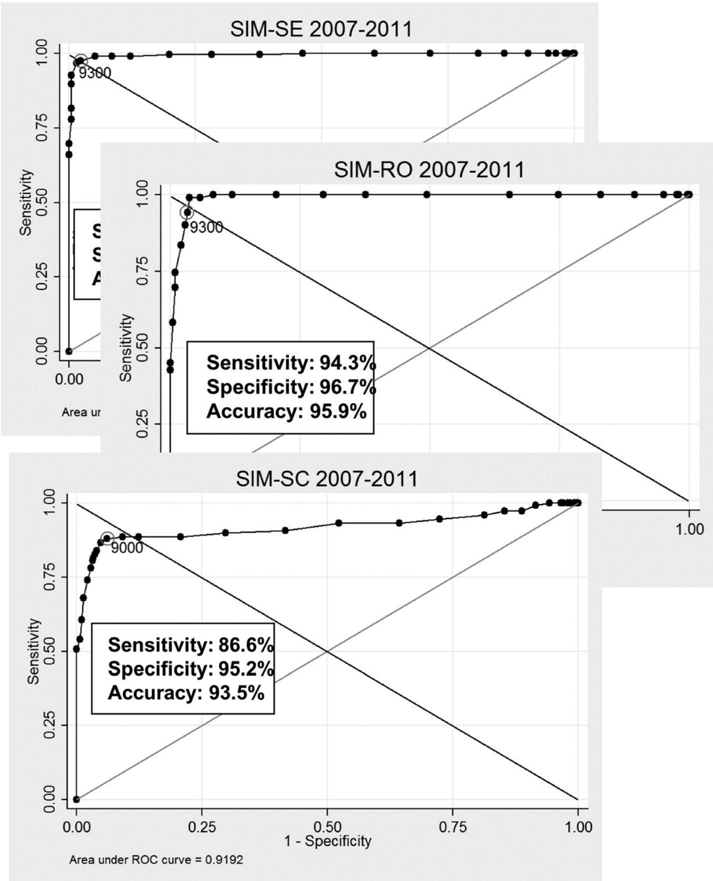
- ✓ Malaria episodes (>5 million records) + mortality + socioeconomic + climate data, 2000 – 2018.
- ✓ Focus on i) data aggregation and ii) epidemic forecasting.

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

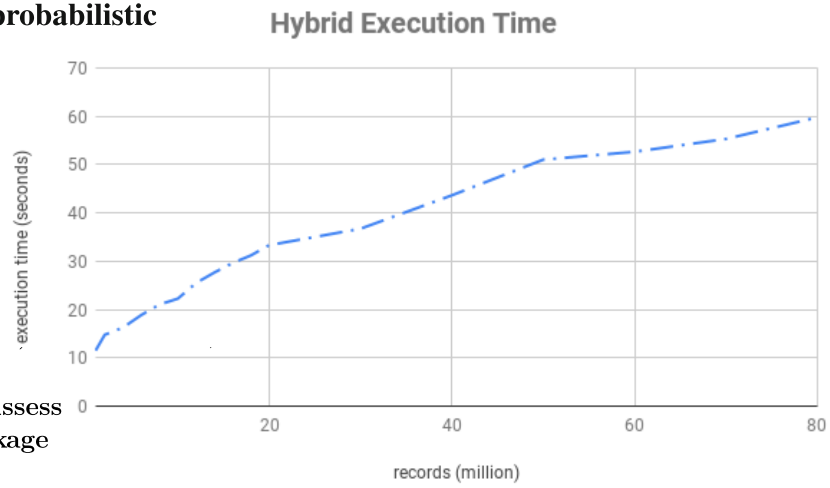
Robespierre Pita, Clícia Pinto, Samila Sena, Rosemeire Fiaccone, Leila Amorim, Sandra Reis, Mauricio L. Barreto, Spiros Denaxas, and Marcos Ennes Barreto



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]

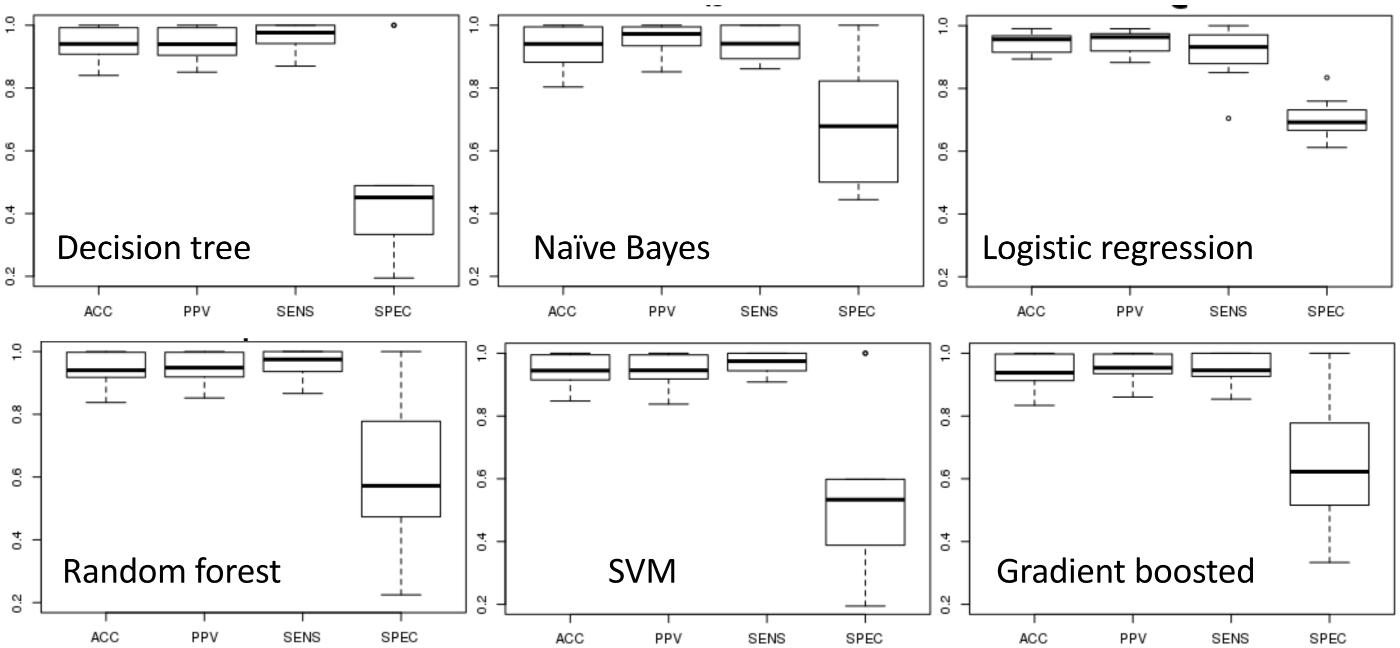
J Supercomput
https://doi.org/10.1007/s11227-018-2328-3



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]



data linkage & analytics

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]

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

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

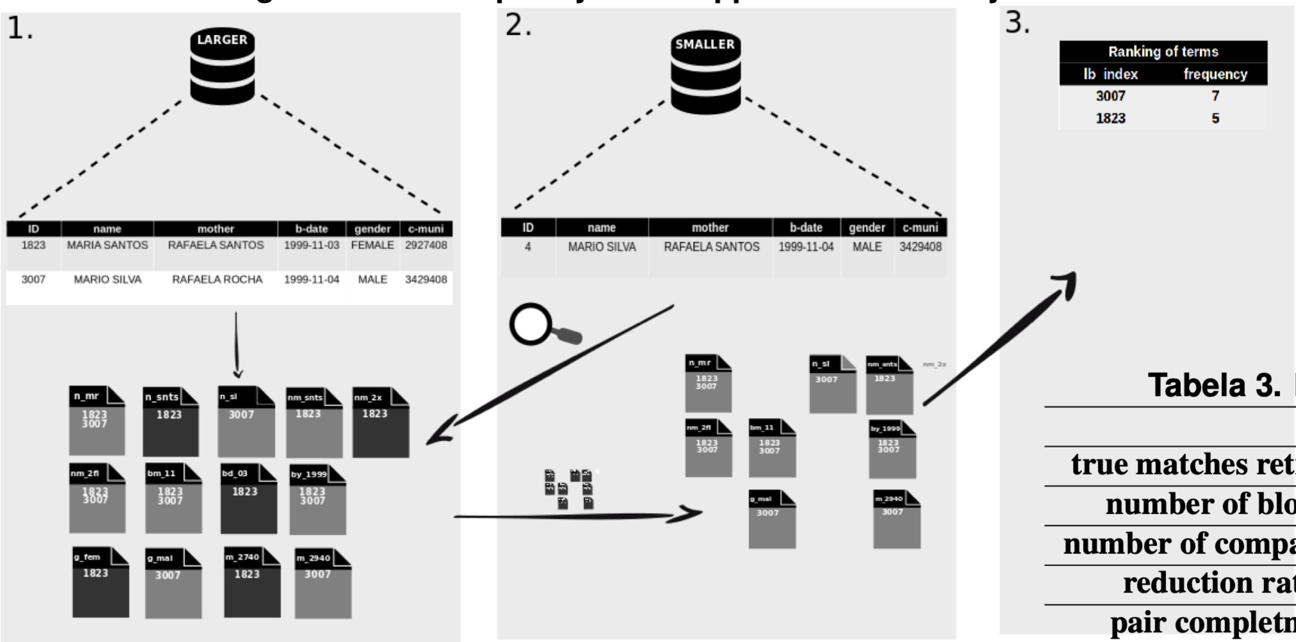


Tabela 3. Results of each indexing techinique used.

	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

data linkage & analytics

✓ **Integrating socioeconomic and healthcare data to combat malaria**

✓ Phase I: November 2016 - October 2018

✓ Focus on i) data aggregation and ii) epidemic forecasting.

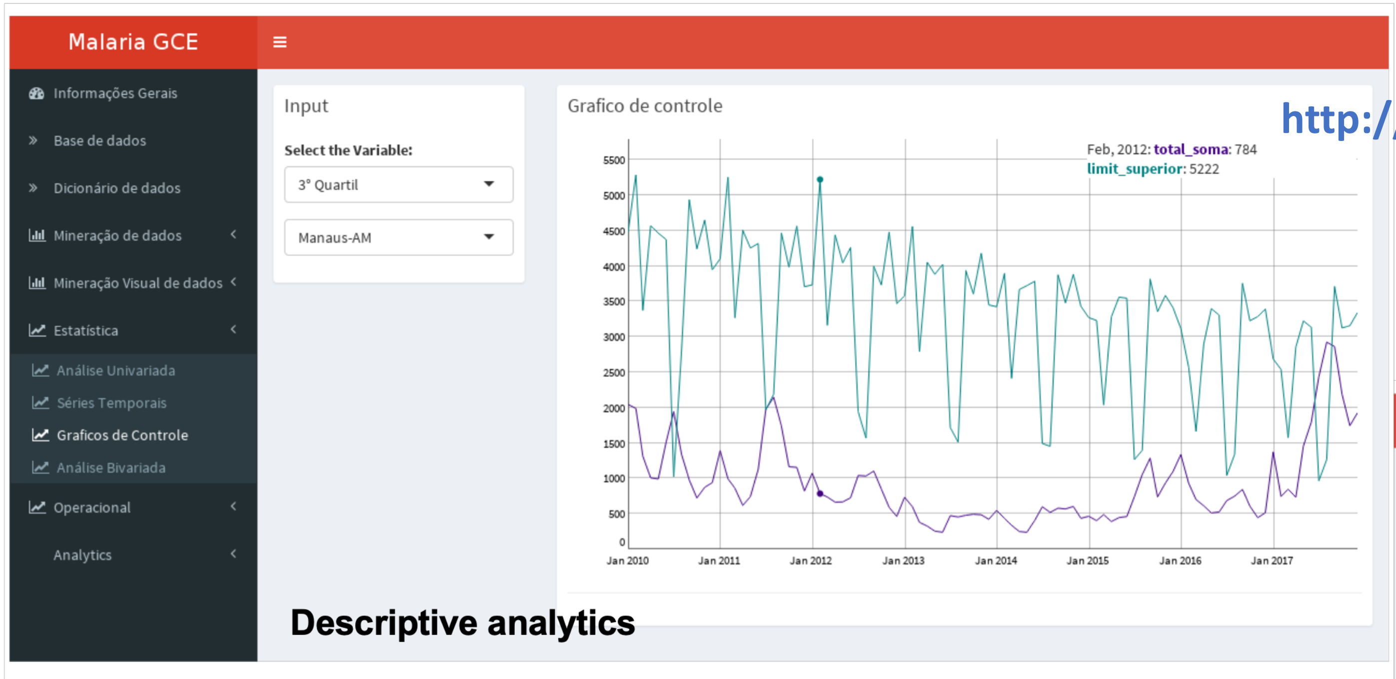
SIVEP
- ✓ Coverage: 2003-2018
- ✓ Records: 5,340,564
- ✓ Attributes: 52

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

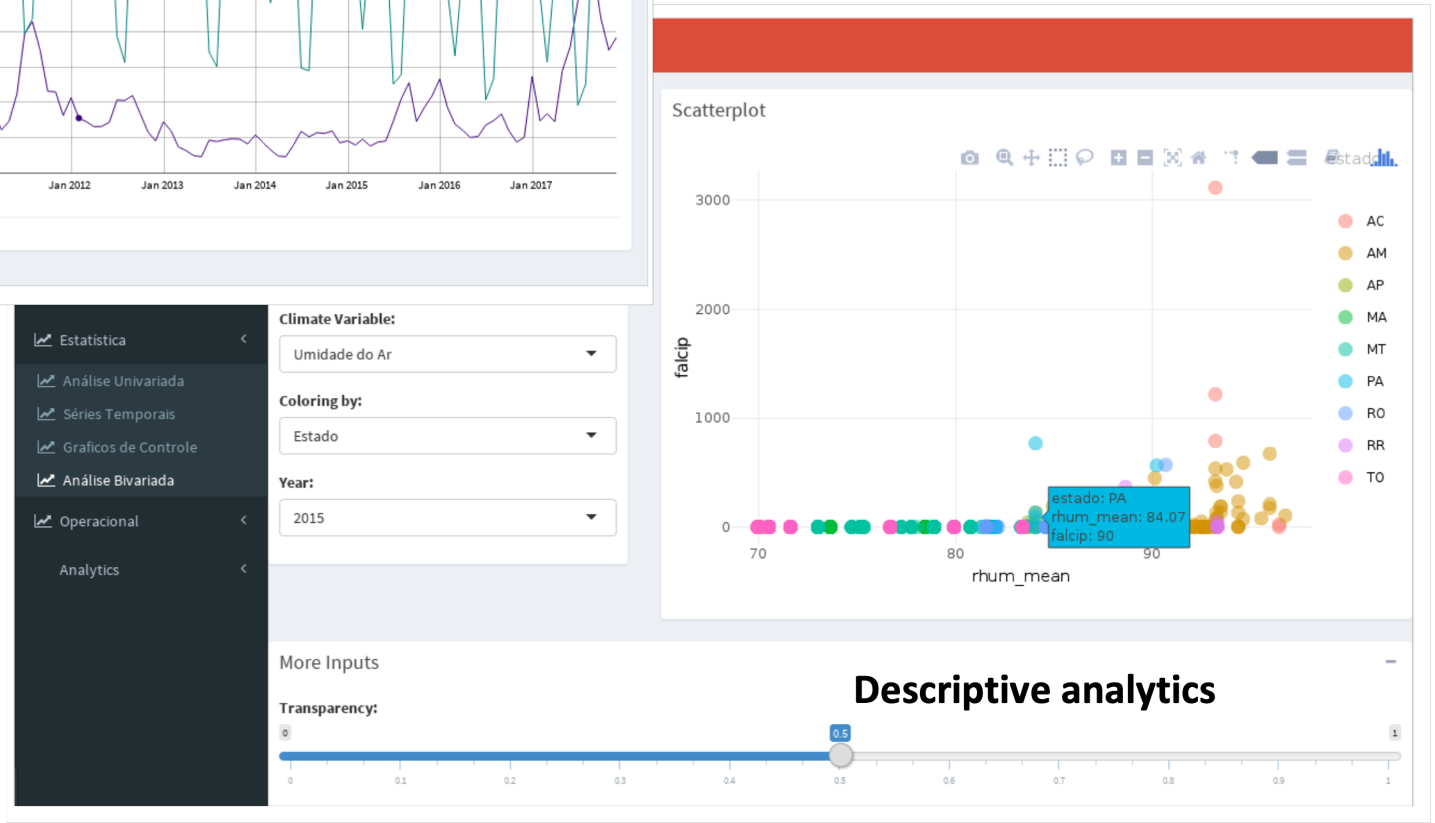
SINAN
- ✓ Coverage: 2003-2018
- ✓ Records: 46,170
- ✓ Attributes: 20

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

UFBA · FIOCRUZ Fundação Oswaldo Cruz · FVS Fundação de Vigilância em Saúde · UCL · Bill & Melinda Gates foundation http://200.128.60.86:3838/shiny_integracao/

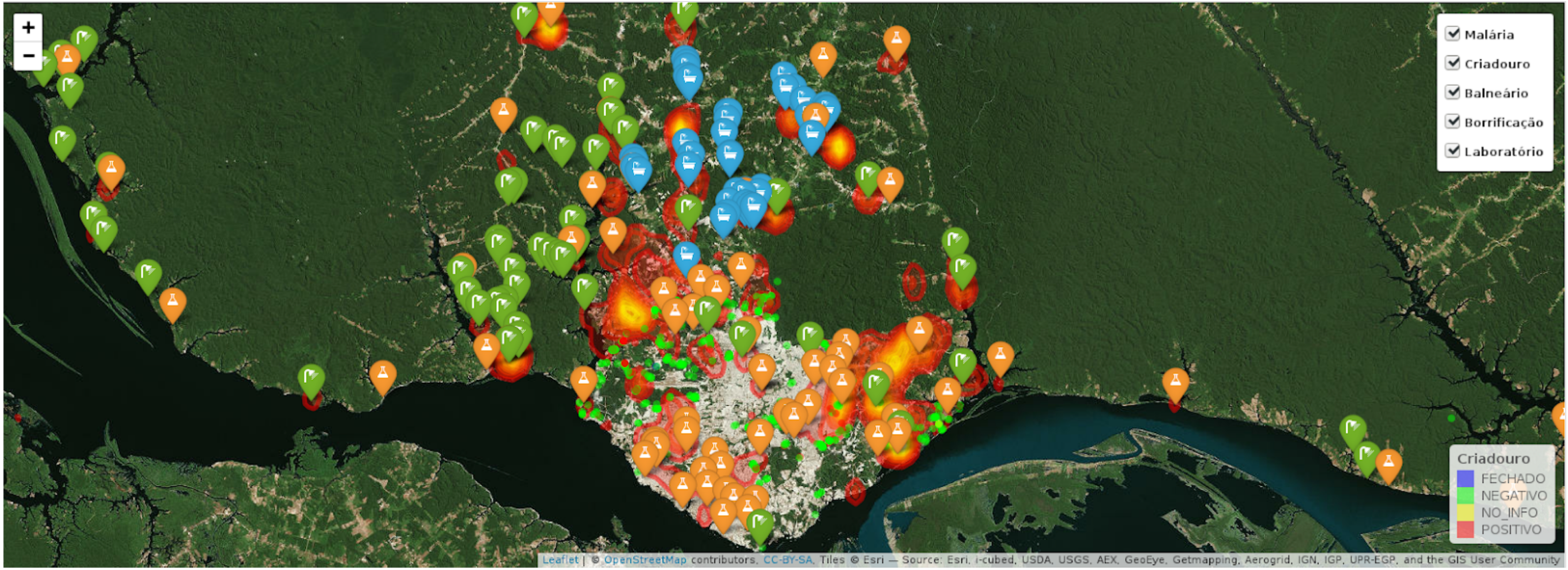


Descriptive analytics



Descriptive analytics

Example: multilayer visual mining



Existing results

MEDTROP
54º CONGRESSO DA SOCIEDADE BRASILEIRA DE MEDICINA TROPICAL

INTEGRAÇÃO E MINERAÇÃO VISUAL DE DADOS PARA ESTUDO DA MALÁRIA NO BRASIL

ALBERTO SIRONI[1], JURACY BERTOLDO JUNIOR[1], MARCOS E. BARRETO[1], VANDERSON SAMPAIO[2], ANDRÉ SIQUEIRA[3]
[1] DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO, UNIVERSIDADE FEDERAL DA BAHIA (SALVADOR, BA), [2] FUNDAÇÃO DE VIGILÂNCIA EM SAÚDE DO AMAZONAS (MANAUS, AM), [3] INSTITUTO NACIONAL DE INFECTOLOGIA EVANDRO CHAGAS (RIO DE JANEIRO, RJ)

International Journal of Population Data Science (2018) 3:3:453

International Journal of Population Data Science

IJPDS International Journal of Population Data Science

Swansea University Prifysgol Abertawe

Journal Website: www.ijpds.org

Linking surveillance and climate data to combat malaria

Sironi, A[1], Barreto, M[1], Bertoldo, J[1], Conceição, D[1], and Sampaio, V[2]

SINAPE 2018

Título

INTERACTIVE DATA VISUALIZATION OF MALARIA USING R SHINY

Autores

André Alves Ferreira Mendes, Rosemeire Leovigildo Fiaccone, Leila Denise Alves Ferreira Amorim, Marcos Ennes Barreto, Juracy Bertoldo Santos Junior, Alberto Sironi, Marcos Aurelio Eustorgio Filho

Resumo

Data visualization consists in representing data in some systematic form including attributes and variables for the unit of information. A simple and quick information can highlight possible errors with data just as it helps uncover interesting trends. There are traditional and new approaches for visualization methods. Data coming from different sources (SIVEP, IBGE, NOAA, MDS) were considered and are potentially useful for the effective understanding of malaria in the Amazon region. Using data of malaria, we illustrated the process of exploratory analysis and traditional visualization tools that can make the evaluation of high dimensional data available and feasible. Exploratory analysis tools (univariate and bivariate) were used to enhance the data visualization techniques. We believe that data integration and the exploration of visualization tools, such as those available using R Shiny, can assist decision making, provide significant contribution for understanding several processes and make massive amounts of available data in several systems become useful.

ROYAL ACADEMY OF ENGINEERING

Frontiers of Engineering for Development symposium:

Engineers as healthcare practitioners

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

CLOSER 2019
9TH INTERNATIONAL CONFERENCE ON CLOUD COMPUTING AND SERVICES SCIENCE
HERAKLION, CRETE - GREECE 2 - 4 MAY, 2019

Current research / projects

Centro de Integração de Dados e Conhecimentos para Saúde
FIOCRUZ

- **The 100 million Brazilian cohort**
  - Principal investigators: Maurício Barreto (FIOCRUZ BA), Gerson Penna (FIOCRUZ DF), Laura Rodrigues (London School of Hygiene and Tropical Medicine), Liam Smeeth (London School of Hygiene and Tropical Medicine).
  - Period: 2015-2019
  - Scope: i) integration of socioeconomic data from Cadastroúnico and Bolsa Família (conditional cash transfer programme) databases to build a huge population-based cohort covering the period 2007-2015. Current cohort size is **114 million records**; ii) design a probabilistic data linkage tool (**AtyImo**) to link this cohort with Public Health databases and generate domain-specific data from several epidemiological studies on HIV, tuberculosis, leprosy etc; iii) propose and validate statistical approaches to probabilistic linkage of huge datasets; iv) promote technology transfer and capacity building on big data integration.
    More information **here**.

Funding:

- **Long-term surveillance platform for Zika virus and microcephaly**
  - Principal investigators: Maurício Barreto (FIOCRUZ/BA), Maria Glória Teixeira (UFBA), Cláudio Henriques Maierovisch (Ministry of Health).
  - Period: 2016-2020
  - Scope: i) Design a cohort based on live births (from SINASC database) from 2001 to 2030; ii) assessment of health and educational outcomes related to Zika virus and microcephaly.
    More information **here**.

Funding:

- **The 100 million Brazilian linked data and datacentre.**
  - Principal investigators: Mauricio Barreto (FIOCRUZ/BA), Laura Rodrigues (London School of Hygiene and Tropical Medicine).
  - Period: 2017-2022
  - Scope: i) link electronic health records from Brazilian governmental databases; ii) build the **CIDACS** datacentre and its public interface.
    More information **here**.

Funding:

Current research / projects

- **Design of a scientific repository (data lake) for big data applications**

Funding:

fapesb

- Principal investigator: Marcos Barreto (UFBA).
- Period: 2016-2019
- Scope: Design and deployment of a data repository (data lake) for big data applications. The first prototype comprises malaria surveillance data to support predictive analytics.

- **Treating heterogeneity and uncertainty in data integration: case study on Brazilian databases.**

Funding:

THE ROYAL SOCIETY

CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

- Principal investigators: Marcos Barreto (UFBA), Spiros Denaxas (Farr Institute of Health Informatics Research).
- Period: 2016-2018
- Scope: i) design and validation of a data integration model and related computing tools addressing heterogeneity, uncertainty and scalability targeted to big data integration; ii) support for some Brazil-UK ongoing projects: the 100 million cohort, the surveillance platform for Zika and microcephaly, and predictive analytics methods applied to Malaria data (Post-doctoral proposal).
  More information **here**.

Current research / projects

- **Integrating socioeconomic and health data to combat malaria.**
  - Principal investigator: Marcos Barreto (UFBA), Spiros Denaxas (Farr Institute of Health Informatics Research).
  - Period: 2016-2018
  - Scope: i) develop a platform to integrate surveillance data from Malaria with other sources (socioeconomic and public health data); ii) design and validate predictive analytics methods to help on Malaria elimination.
    More information **here**.

Funding:

BILL & MELINDA GATES foundation

- **Early childhood development friendly index: assessing the enabling environment for Nurturing Care.**
  - Principal investigator: Muriel Gubert (UnB).
  - Team: Marcos Barreto (UFBA), Gabriela Buccini (Yale School of Public Health), Rafael Perez-Escamilla (Yale School Of Public Health), Sonia Isoyama Venancio (Health Institute of São Paulo)
  - Period: 2018-2020
  - Scope: This project aims to develop an ECD (Early Childhood Development) friendly index (ECD-FI), based on a core set of evidence-based Nurturing Care indicators, to assess the enabling environment and promote ECD at the municipality level by monitoring and identifying opportunities to scale up ECD programs locally. More information **here**.

Funding:

BILL & MELINDA GATES foundation

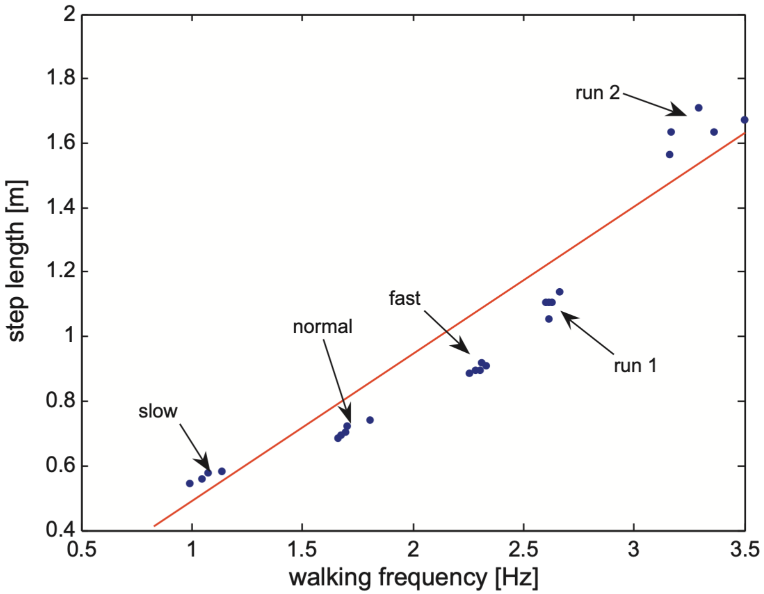
CNPq Conselho Nacional de Desenvolvimento Científico e Tecnológico

Current research / projects

- **Standardisation of wearable-based algorithms for healthcare applications in developing countries.**
  - Principal investigator: Alan Godfrey (Northumbria University, Newcastle-upon-Tyne, UK).
  - Team: Rodrigo Vitorio (UNESP), Marcos Barreto (UFBA), Azad Hussain (University of Birmingham), Clara Aranda-Jan (University College London)
  - Period: 2018-2019
  - Scope: This proposal aims tol develop a novel standardised framework to better inform algorithms for a more harmonised gait assessment in Parkinson's disease (PD), particularly for developing countries where guidance is lacking. This project will lead to the 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. This will better guide PD assessment for sustainable health, promoting and encouraging low-cost wearables as routine diagnostics in developing countries. This framework will also be adapted to the needs of those in developed regions.

Funding:

ROYAL ACADEMY OF ENGINEERING



(a) Walking frequency versus step length

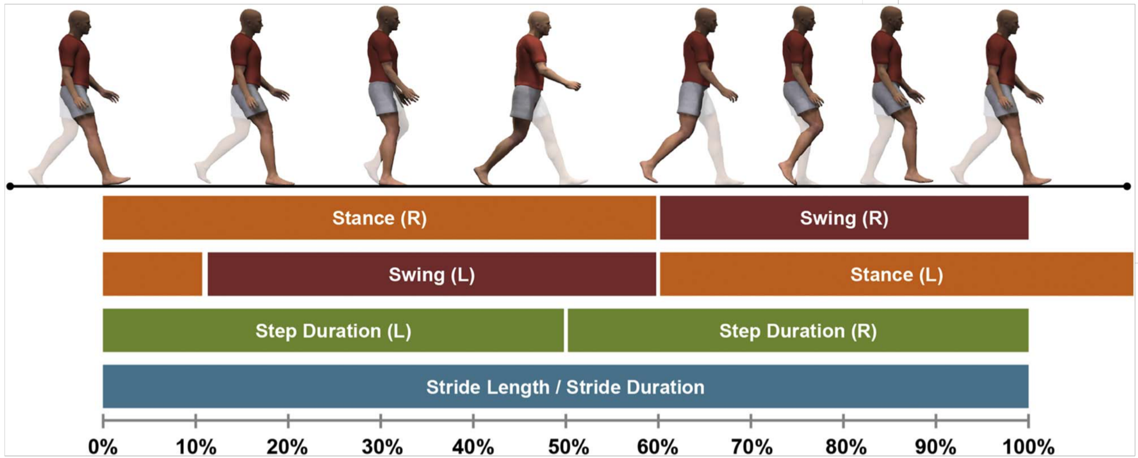


Fig. 1. Schematic of the human gait cycle and the spatiotemporal parameters validated in this study. Specifically, we validated the IMU system's ability to measure the stance percent, swing percent, stride duration (gait cycle time), stride length, and step duration in addition to the speed and cadence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

Current research

- **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

OBRIGADO!

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