Eulac PerMed 2019
Montevideo - Uruguay

Experiences on Data Science in Health

Fabio Porto (fporto@lncc.br),
LNCC – MCTIC
DEXL (http://dexl.lncc.br)

Laboratório Nacional de Computação Científica

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES

BRASIL GOVERNO FEDERAL

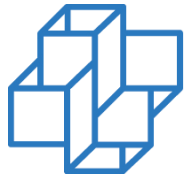
Laboratório Nacional de Computação Científica

LNCC - DEXL

Eulac PerMed 2019

DEXL
DATA EXTREME LAB

National Laboratory of Scientific Computing



- Research unit of Ministry of Science Technology, Innovation and Communication, Brazil
- Graduate Course in Computational Modeling
- Coordinator of the SINAPAD
- INCT-MACC
- INCT-CID
- Thematic Laboratories
  - LabInfo – Bioinformatics
  - Hemolab – Cardio-Vascular System
  - DEXL – Data Extreme Lab

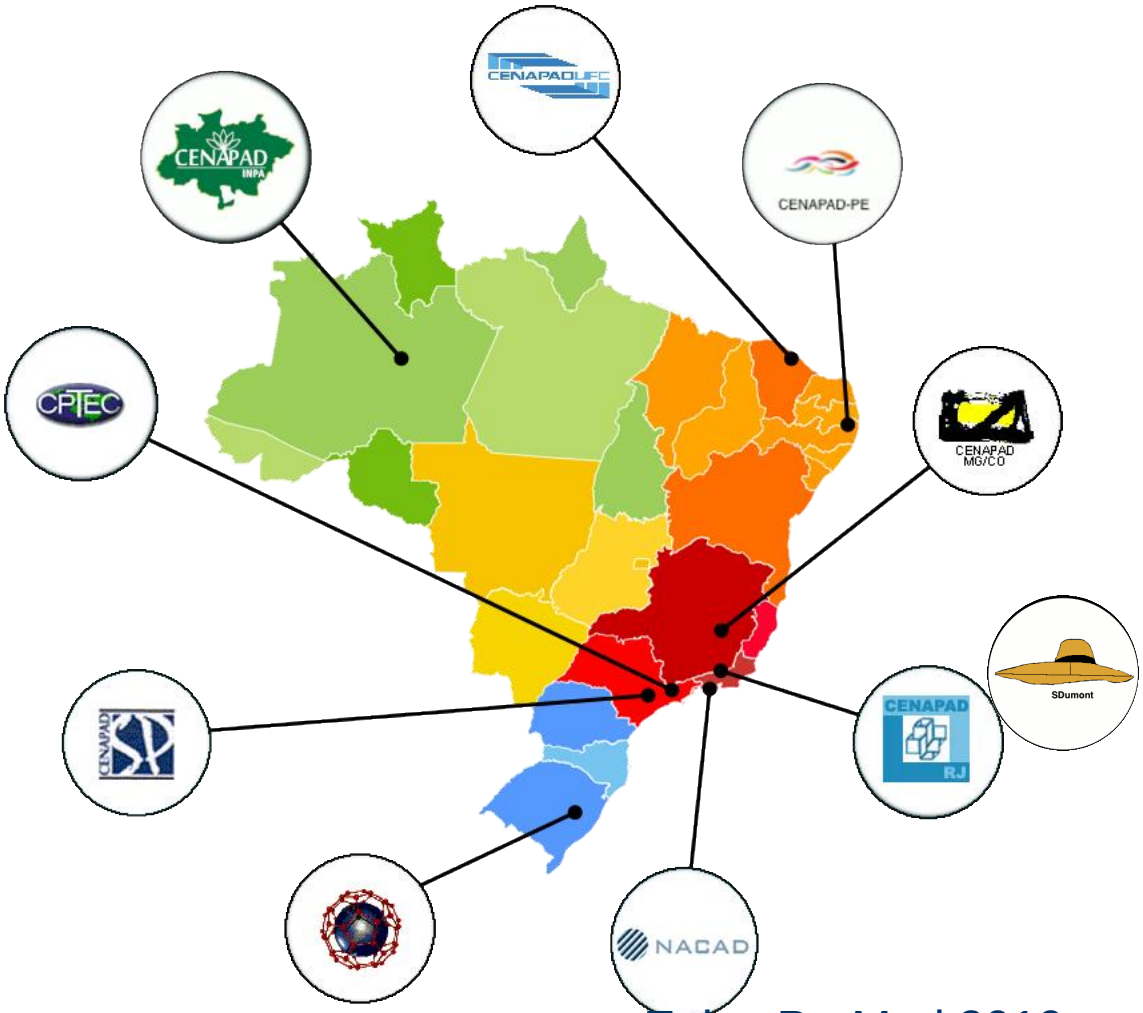
DEXL
DATA EXTREME LAB

Petropolis, Rio de Janeiro, Brazil

Eulac PerMed 2019

DEXL
DATA EXTREME LAB

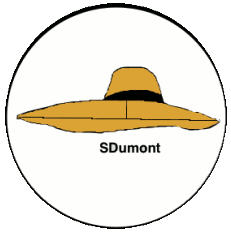
SINAPAD: National System of HPC



Eulac PerMed 2019

DEXL
DATA EXTREME LAB

Santos Dumont – Super Computer



Total peak capacity- 1,1 petaflops

- 18.144 CPU cores on 756 computing node (24 cores per node);

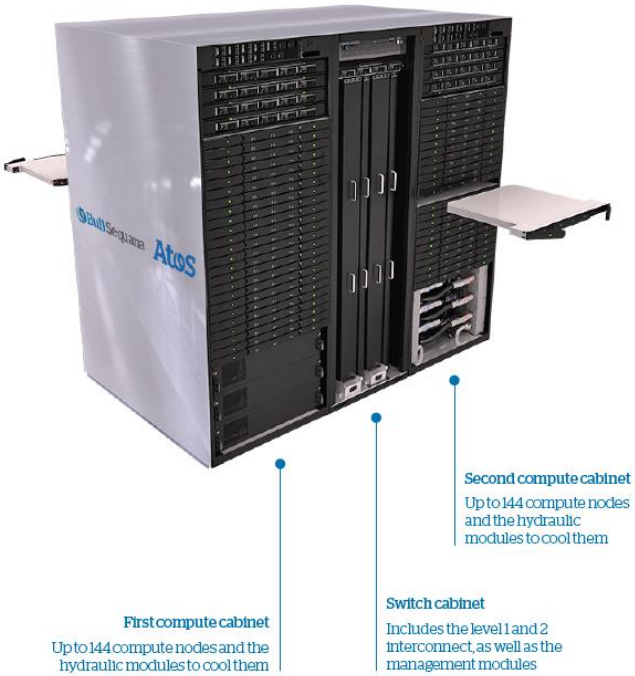
Nodes with GPUs NVIDIA K40 and Xeon PHI

- fatnode with 6TB RAM and 16 CPU Intel Ivy (15 cores per CPU))

DEXL
DATA EXTREME LAB

Santos Dumont Expansion – Nov/2019 (top 500)

- 2 Sequana X1000 Modules (Atos-BULL
  - 82 Blades X1120
  - 12 Blades X1120
  - 94 Blades X1125
  - Total top capacity 4 Pflops
- Total peak capacity 5.1 pflops



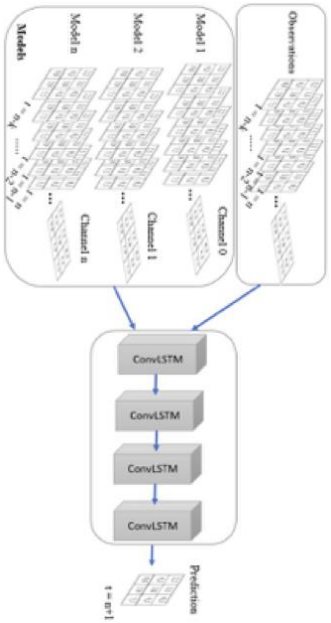
193	Laboratório Nacional de Computação Científica Brazil	**Santos Dumont (SDumont)** - Bull Sequana X1000, Xeon Gold 6252 24C 2.1GHz, Mellanox InfiniBand EDR, NVIDIA Tesla V100 SXM2 Atos	33,856	1,849.0	2,727.0

DEXL DATA EXTREME LAB

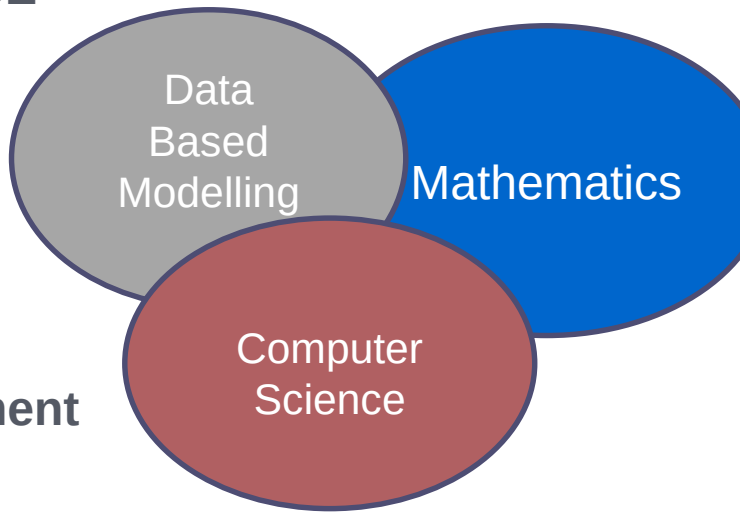
DEXL Overview

ARTIFICIAL INTELLIGENCE

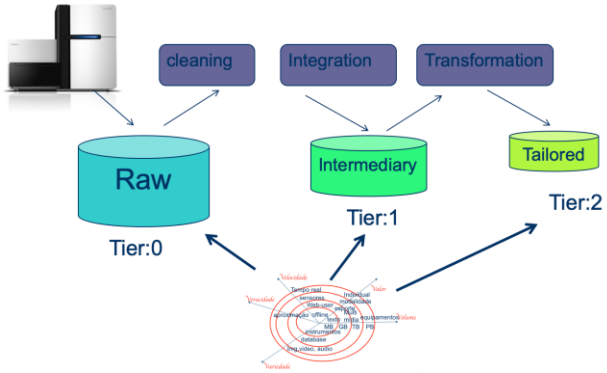
DATA SCIENCE



DEXL Lab



BIG DATA Processing &Management



DEXL
DATA EXTREME LAB

Introduction

- Modern Information systems comprise a complex set of methods to make sense of the observed phenomenon
  - Data, Models, UIs
- Data comes in different forms: signals, measurements, images, simulation data etc..
- Knowledge expressed through different approaches: knowledge bases, models (first principles, Machine Learning,..), web, texts, pdf files,...
- Different tools to explore knowledge sources

DEXL
DATA EXTREME LAB

Objective

- In this talk I will convey the message that personalized follow-up requires a holistic view on heterogeneous knowledge sources.

  Our approach: Methods + tools

Outline

- Knowledge Base Introduction
- THE SAHA SYSTEM
- THE DATA SCIENCE SUITE
- THE SAVIME SYSTEM
- Final Remarks

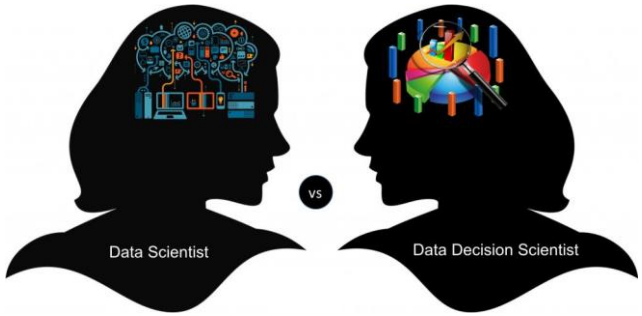
KNOWLEDGE BASE CONSTRUCTION

DEXL
DATA EXTREME LAB

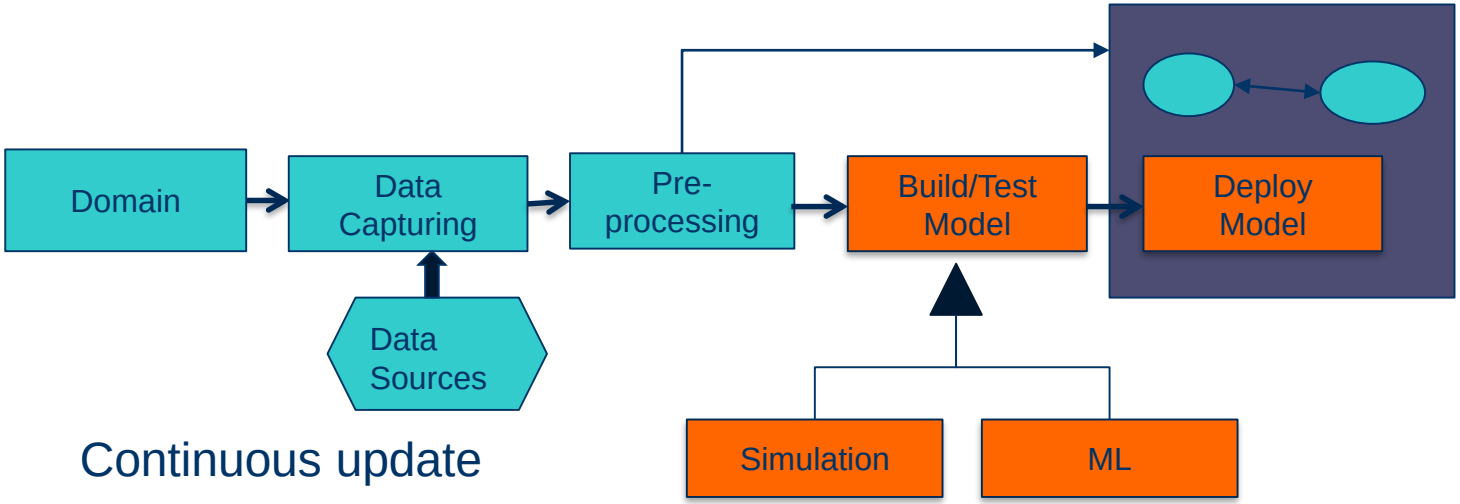
Knowledge Bases

- KB for specific Domains
- An extension to databases approach with emphasis on relationships and inferencing methods
- A graph-based representation comprising entities, their relationships, classes of entities all represented at a same logical level
- Accommodates different data types
- Languages for representation, navigation and reasoning

Knowledge Base Construction



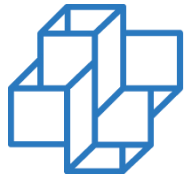
KnowledgeBase



Continuous update

DEXL
DATA EXTREME LAB

KB – Construction Process Overview



Data Management

- Data Capturing
- Data Cleaning
- Normalization, Uniformization
- Enrichment (external sources)
- KB Population

Model Management + KB update

- Data Correlation
- Trainining Predictive Model
- Relationship inference
- Simulation Model

KB consumption

- Data set export
- Query/ Search
- Visualization

DEXL
DATA EXTREME LAB

System developed by DEXL to the Brazilian Olympic Committee to improve athletes performance

SAHA – ATHLETES HOLISTIC FOLLOW-UP SYSTEM

http://dexlservice.lncc.br/saha

Fabio Porto, Ana Maria Moura, F. C da Silva et al., A Metaphoric Trajectory Data Warehouse for Olympic Athlete follow-up, Concurrency and Computation: Practice and Experience, 24(23), 2012.

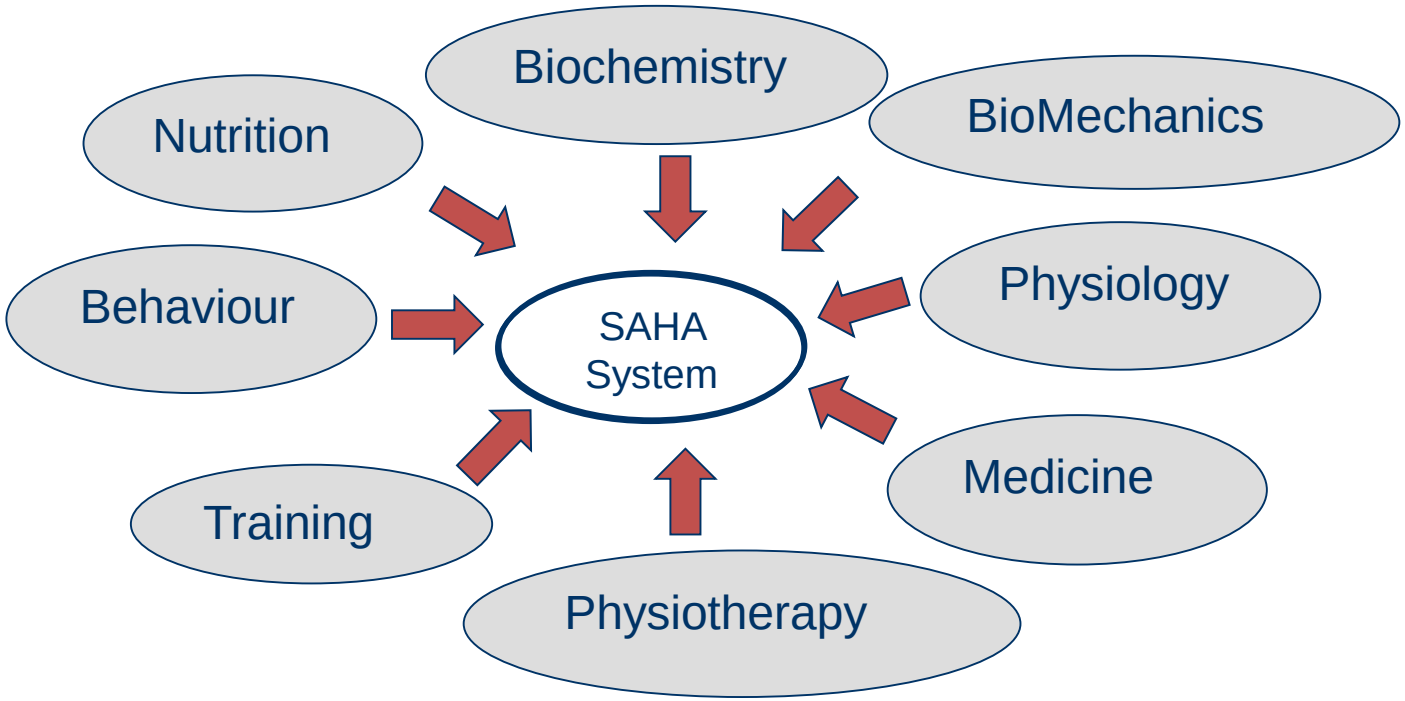
Eulac PerMed 2019

DEXL
DATA EXTREME LAB

Challenges

- Integrate observations from different disciplines
- Extensible to any new quantifiable variable
- Integrate with athlete/patient filled questionnaires
- Common analytical view fostering longitudinal interpretation
- Focus on individual data

Multidisciplinary



A. Bassini, L.C. Cameron, Sportomics: Building a new concept in metabolic studies and exercise science, Biochem BioPhys Res Comm, 2014, 445 (4).

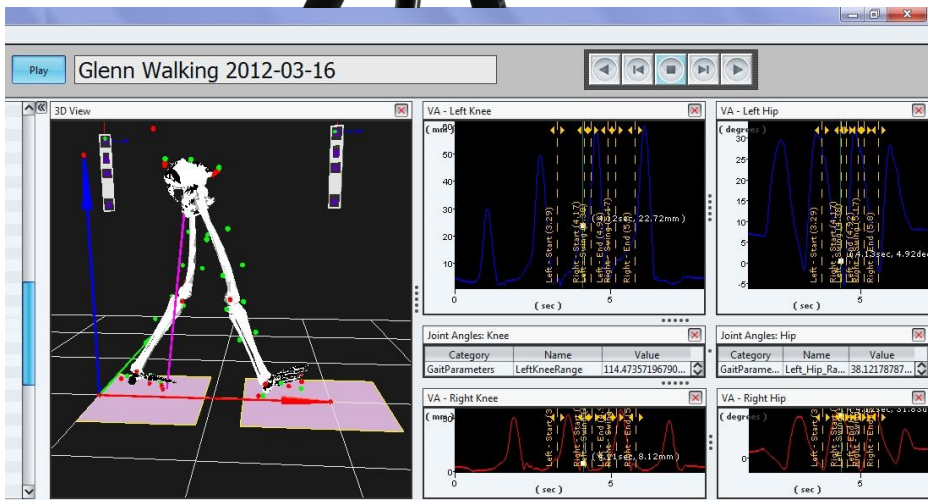
DEXL
DATA EXTREME LAB

HETEROGENEOUS OBSERVATIONS EXTRACTION AND LOAD

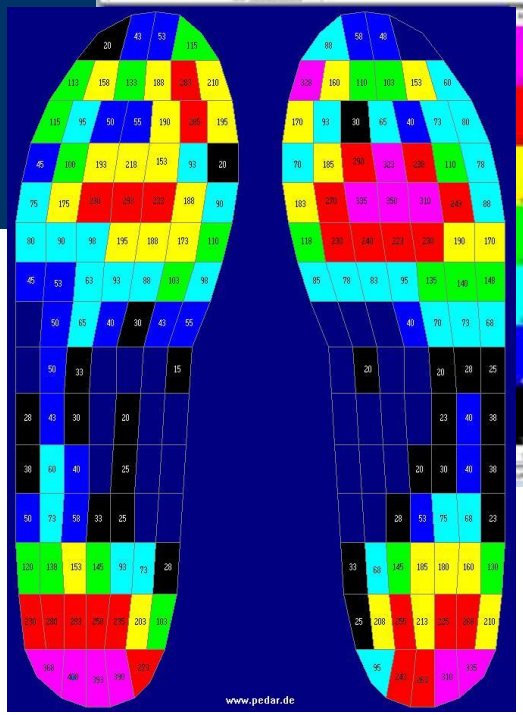
Eulac PerMed 2019

DEXL
DATA EXTREME LAB

codamotion


Play
Glenn Walking 2012-03-16
3D View
VA - Left Knee
VA - Left Hip
Joint Angles: Knee
Category
Name
Value
GaitParameters
LeftKneeRange
Joint Angles: Hip
GaitParame...
Left_Hip_Ra...
VA - Right Knee
VA - Right Hip
(sec)

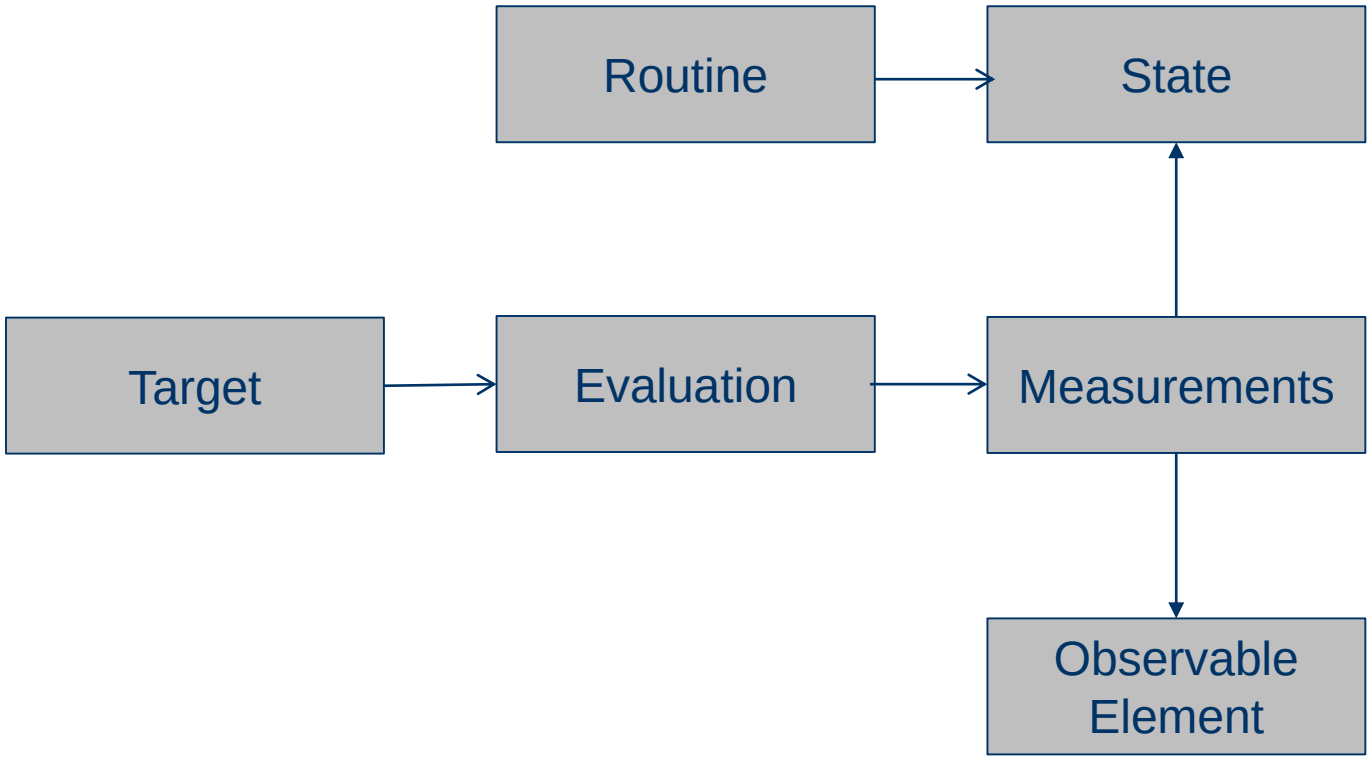
novel.de



Blood Collection – Blood Count data

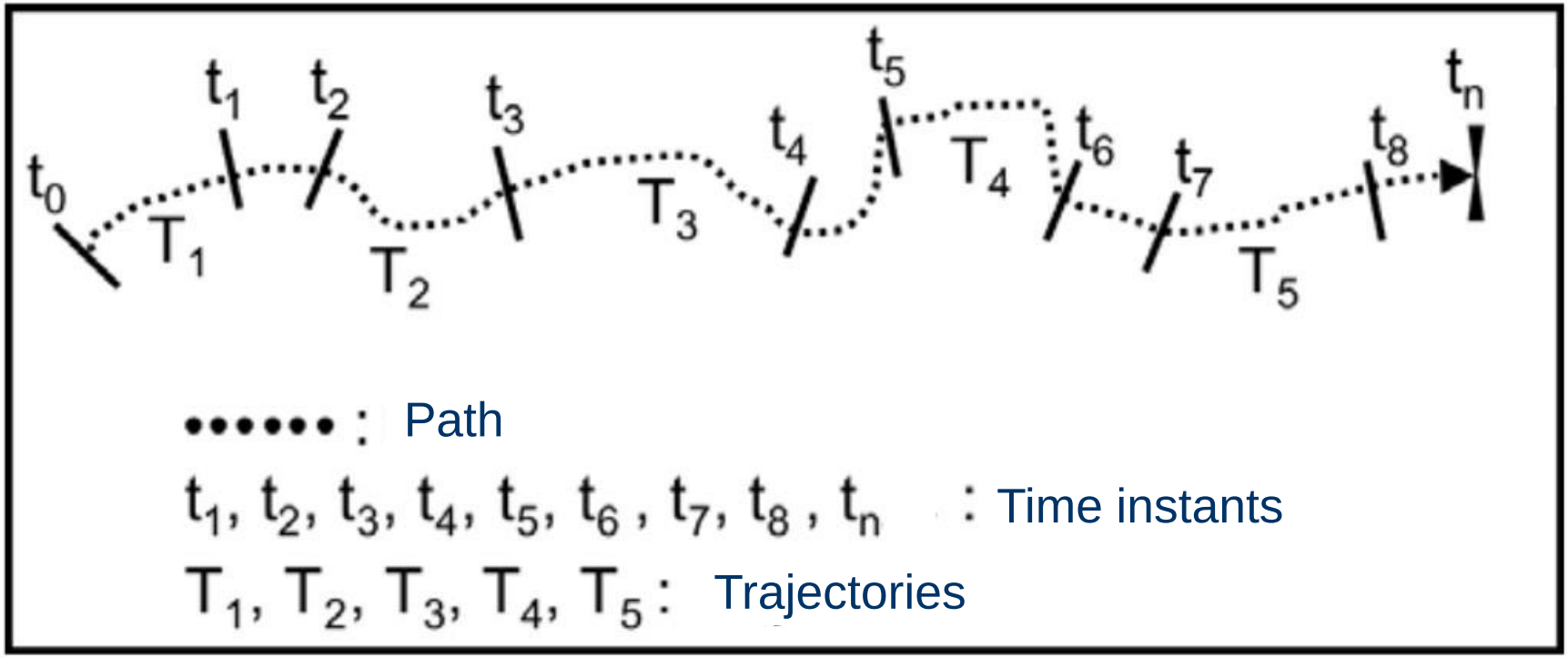
DEXL
DATA EXTREME LAB

Data Schema – High Level/ Simplified View



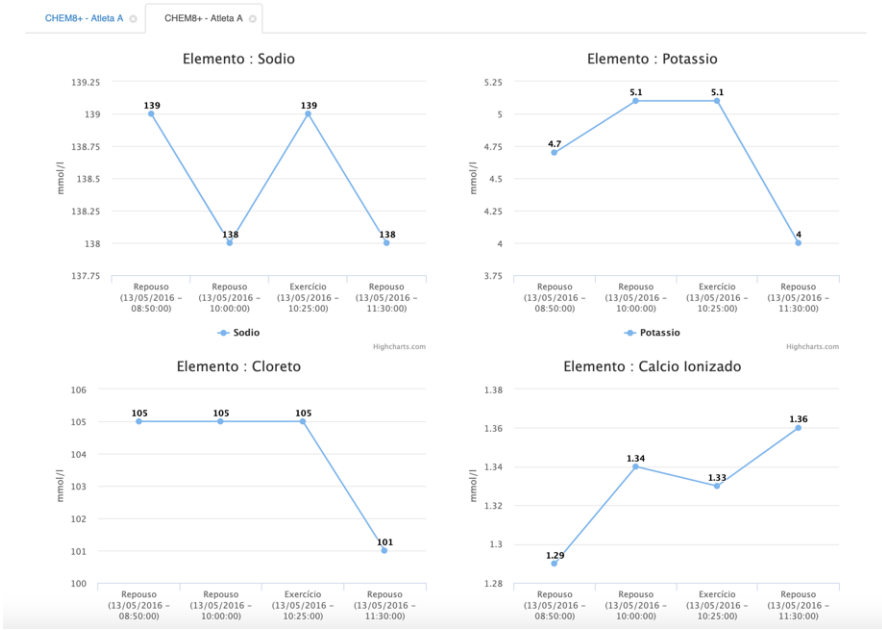
DEXL
DATA EXTREME LAB

Observable elements follow-up as trajectories



DEXL
DATA EXTREME LAB

Longitudinal analysis on historical measurements



http://dexlservice.lncc.br/saha

Eulac PerMed 2019

DEXL
DATA EXTREME LAB

Longitudinal observations of a OE

Eulac PerMed 2019

DEXL
DATA EXTREME LAB

From Observations to Evidences proving hypothesis

Evaluation target: To reduce running time

Evaluation Target

Evidence

Hypothesis

Suggestions

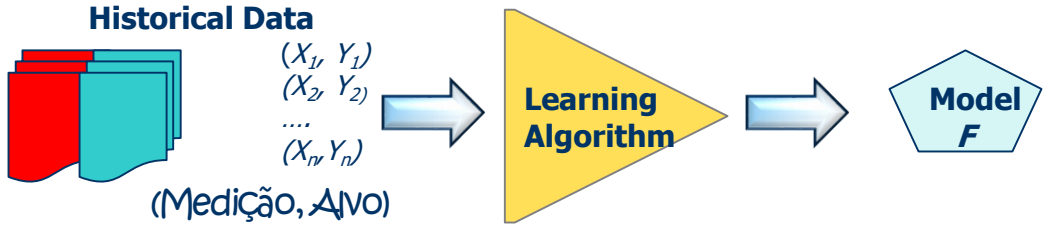
ESPORTE	OBJETIVO	PRAZO	#	DEPTO.	EVIDÊNCIAS	HIPÓTESES		SUGESTÕES
MARATONA	REDUÇÃO DO TEMPO EM 2 MINUTOS	1 ANO	1	BIOQUIMICA	Análise do nível de CK, GamaGT	Aumentar Treinamento		Implementação de práticas nutricionais suplementares
					Queda da concentração de bicarbonato	Aumentar reserva de bicarbonato		Aumentar treinamento
					Níveis de Creatinina	Melhorar hidratação		
					Níveis de Glicose, Ureia e Creatinina	Rever status nutricional		
					Baixa reserva de glicogênio	Dieta inadequada ou treinamento leve		
			2	NUTRIÇÃO	Baixa ingestão de água	Falta de energia		
					Baixa reposição de energia			
					Baixa ingestão de carboidratos			
					Consumo baixo de vitamina D			
			3	FISIOLOGIA	Análise da composição corporal	Excesso de Gordura		
					Análise do consumo máximo de oxigênio	Baixo consumo de VO2		
						Baixo Limiar ventilatório 2		
			4	BIOMECÂNICA	Articulação joelho estável	Apto para mais esforço		
						Boa habilidade motora		
						Domínio de força		
			5	COMPORTAMENTO	Índice de auto eficácia	Alta eficácia e Auto Confiança acima da média		
					Índice de Stress/Recuperação	Leve Exaustão emocional		
						Preocupação com lesões		
			6	ANÁLISE DO DESEMP. E SUP. AO TREINO				

Eulac PerMed 2019

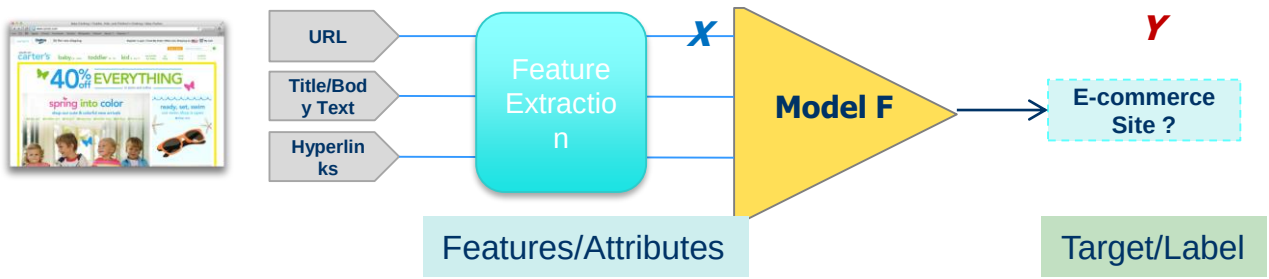
DEXL DATA EXTREME LAB

Supervised Learning

- **Training:** Given training examples $\{(X_i, Y_i)\}$ where X_i is the feature vector and Y_i the target variable, learn a function F to best fit the training data (i.e., $Y_i \approx F(X_i)$ for all i)



- **Prediction:** Given a new sample X with unknown Y, predict Y using $F(X)$



- **Inverted Problem:** Given a Y determine relevant X using $F^{-1}(Y)$

DEXL
DATA EXTREME LAB

Project Status

- SAHA systems entirely functional
- Implementation focused in:
  - Hosting data from different sources
  - Offering an integrated and holistic view on targets
  - Some fixed plots that explore statistical and longitudinal aspects of measurements
- The AI layer to be integrated

Http:// dss.dexl.lncc.br

DATA SCIENCE SUITE (DSS)

Rafael S. Pereira, Fabio Porto, SBBD2019, Demo, Fortaleza, Brazil

Eulac PerMed 2019

DEXL
DATA EXTREME LAB

Objective

- To provide a suite of data science services
- From a data exploratory Perspective to a ready to use ML based predictions
- To be extensible to new features (still under development)

Current Services

TextAnalysis

Application which lets the user explore the contents of a PDF file and compare two different PDFS for similar subjects

DataExploration

Application which lets the user do data exploration on a tabular dataset with a interface

Plant_Classification

Application to classify plant images into possible species and classify their health status

GraphAnalysis

A Application that receives the adjacency matrix of a graph in a tabular form and lets the user see many metrics of this graph

SentimentAnalysis

Application which lets the user explore the contents of a PDF file for different sentiments

Deep_Learning_Image_Predictor

Application which takes as a input a image and classifies it to the 5 most possible labels, expects a keras trained model

TimeSeries

A Application that receives a time series in a vector form inside a tabular file and lets the user analyze the series

MachineLearning

A Application that receives a Tabular file and lets the user test machine learning models to predict the variables

ObjectDetection

A application that receives a image and detects all objects trained on the coco dataset

Eulac PerMed 2019

DEXL
DATA EXTREME LAB

Childbirth per district with

Exploratory Data Analysis

Insira o arquivo a ser analisado.csv

Browse... Atendimentos Hospitalares por Município.csv

Upload complete

- [x] Show Summary
- [ ] Show correlation matrix
- [ ] Show the relative amount of missing data
- [ ] Show the functional dependency(uses x as variable to predict)
- [ ] Show The minimum set of aproximate functional dependency(uses x as variable to predict)
- [ ] Generate best Graph
- [ ] Show histogram of x

X

Cirúrgico

Y

Obstétrico

Color

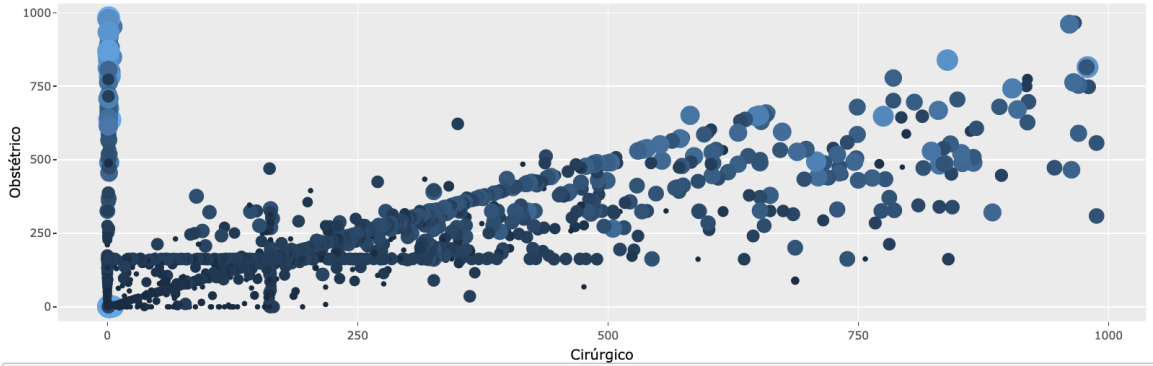
Neonatal

Size

Neonatal

Completar Dataset | Download do dataset completado

Data Exploration | PCA Visualization | Correlation Matrix Image | 3D plots




```
 Município            UF            Cirúrgico       Obstétrico
Length:5570        Length:5570        Min.   :  0.0   Min.   :  0.0
Class :character   Class :character   1st Qu.:  0.0   1st Qu.:  0.0
Mode  :character   Mode  :character   Median :147.0   Median :162.0
                                      Mean   :125.1   Mean   :129.7
                                      3rd Qu.:163.0   3rd Qu.:163.0
                                      Max.   :988.0   Max.   :982.0

   Neonatal
Min.   :  0.0
1st Qu.:  0.0
Median :  0.0
Mean   : 81.3
3rd Qu.:162.0
Max.   :997.0
```

Eulac PerMed 2019

DEXL
DATA EXTREME LAB

Childbirth normal vs cesarean

Exploratory Data Analysis

Insira o arquivo a ser analisado.csv

Browse... Atendimentos Hospitalares por Município.csv

Upload complete

- [x] Show Summary
- [ ] Show correlation matrix
- [ ] Show the relative amount of missing data
- [ ] Show the functional dependency(uses x as variable to predict)
- [ ] Show The minimum set of aproximate functional dependency(uses x as variable to predict)
- [x] Generate best Graph

Termos maximos da expansão

1 4 20

1 3 5 7 9 11 13 15 17 19 20

Numero de modelos a ser plotado

1 2 12

1 3 5 7 9 11 12

Metricas

R squared

- [ ] Show histogram of x

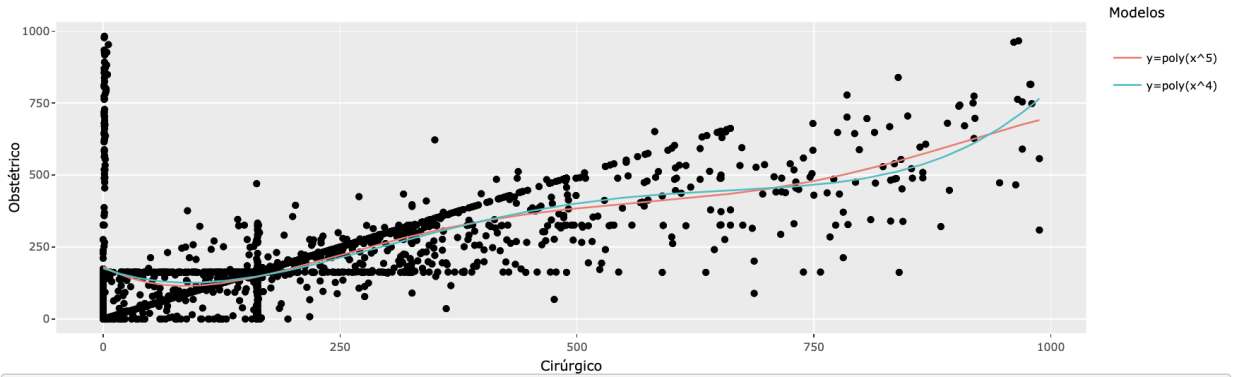
X

Cirúrgico

Y

Obstétrico

Data Exploration | PCA Visualization | Correlation Matrix Image | 3D plots




```
 Município           UF           Cirúrgico      Obstétrico
Length:5570        Length:5570        Min.   :  0.0   Min.   :  0.0
Class :character   Class :character   1st Qu.:  0.0   1st Qu.:  0.0
Mode  :character   Mode  :character   Median :147.0   Median :162.0
                                      Mean   :125.1   Mean   :129.7
                                      3rd Qu.:163.0   3rd Qu.:163.0
                                      Max.   :988.0   Max.   :982.0

   Neonatal
Min.   :  0.0
1st Qu.:  0.0
Median :  0.0
Mean   : 81.3
3rd Qu.:162.0
Max.   :997.0
```

Eulac PerMed 2019

DEXL DATA EXTREME LAB

Prediction: healthy vs not health

Not Healthy

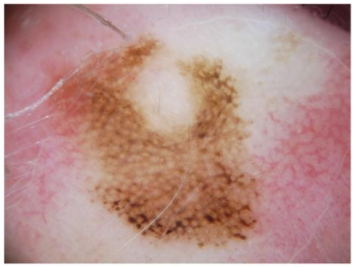
Deep Learning Predictor

Selecione a imagem
Browse... ISIC_0001132.jpg
Upload complete

Selecione o modelo
Browse... cancer.model
Upload complete

Selecione o arquivo de classes csv
Browse... cancerpickle.csv
Upload complete

numero de classes preditas
5



```
     class probability
1   Doente  0.98948050
0 Saudavel  0.01051945
```

Healthy

Deep Learning Predictor

Selecione a imagem
Browse... ISIC_0000000.jpg
Upload complete

Selecione o modelo
Browse... cancer.model
Upload complete

Selecione o arquivo de classes csv
Browse... cancerpickle.csv
Upload complete

numero de classes preditas
5

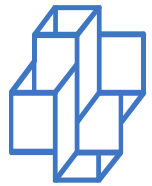
```
     class probability
0 Saudavel   0.8937429
1   Doente   0.1062572
```

Eulac PerMed 2019

DATA EXTREME LAB

Comments

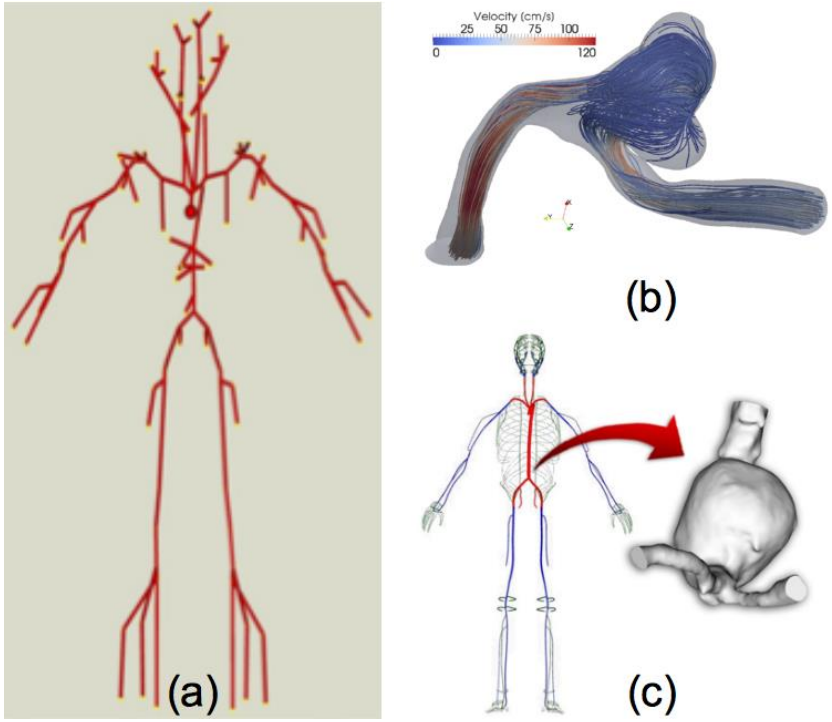
- DSS is available for use through web interface
- Aims at supporting experienced and novice data scientists in extracting knowledge from data
- Is in ongoing development
  - Rafael Silva Pereira, LNCC Msc student



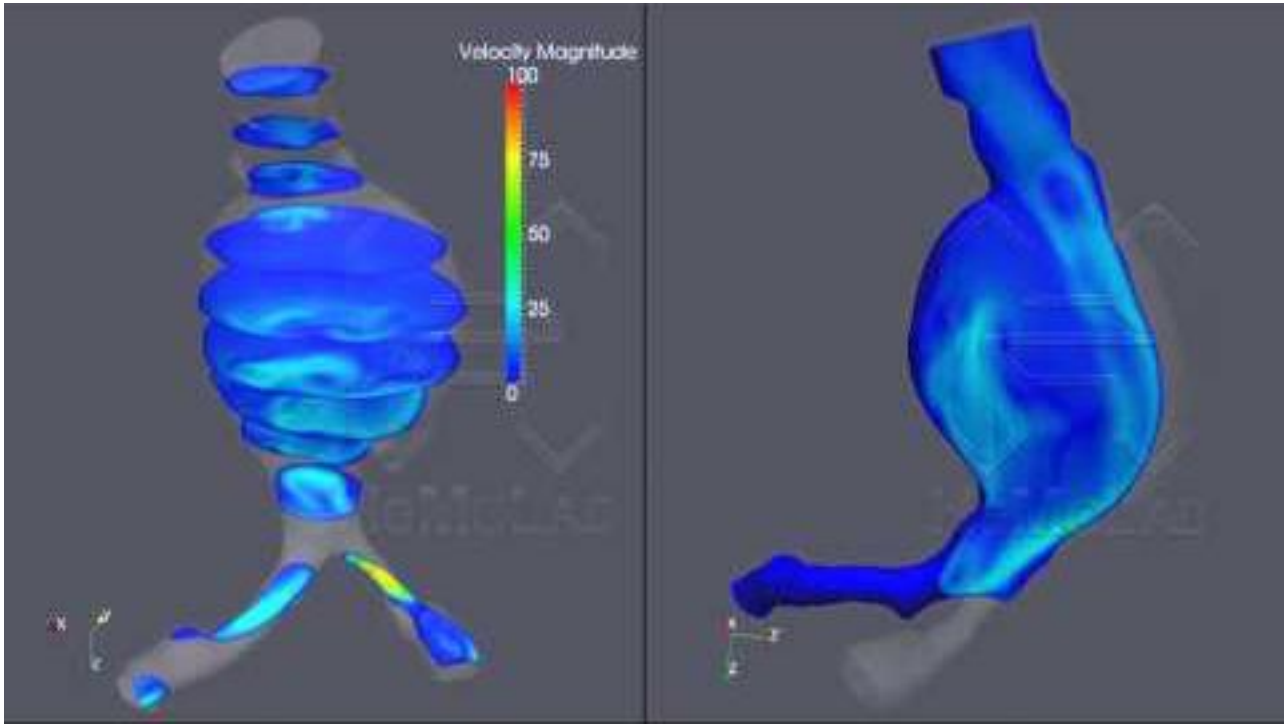
SAVIME Simulation Data Analysis & Visualization

DEXL
DATA EXTREME LAB

Simulations: CardioVascular System – Hemolab LNCC

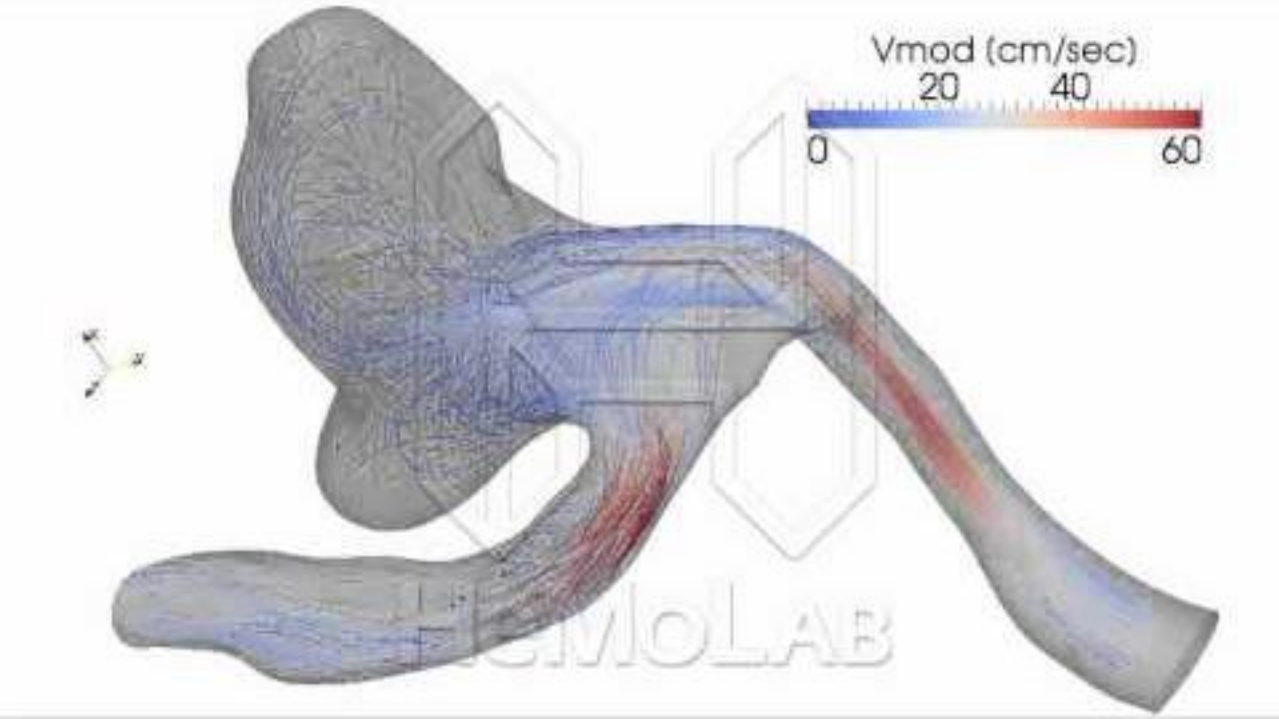


DEXL
DATA EXTREME LAB



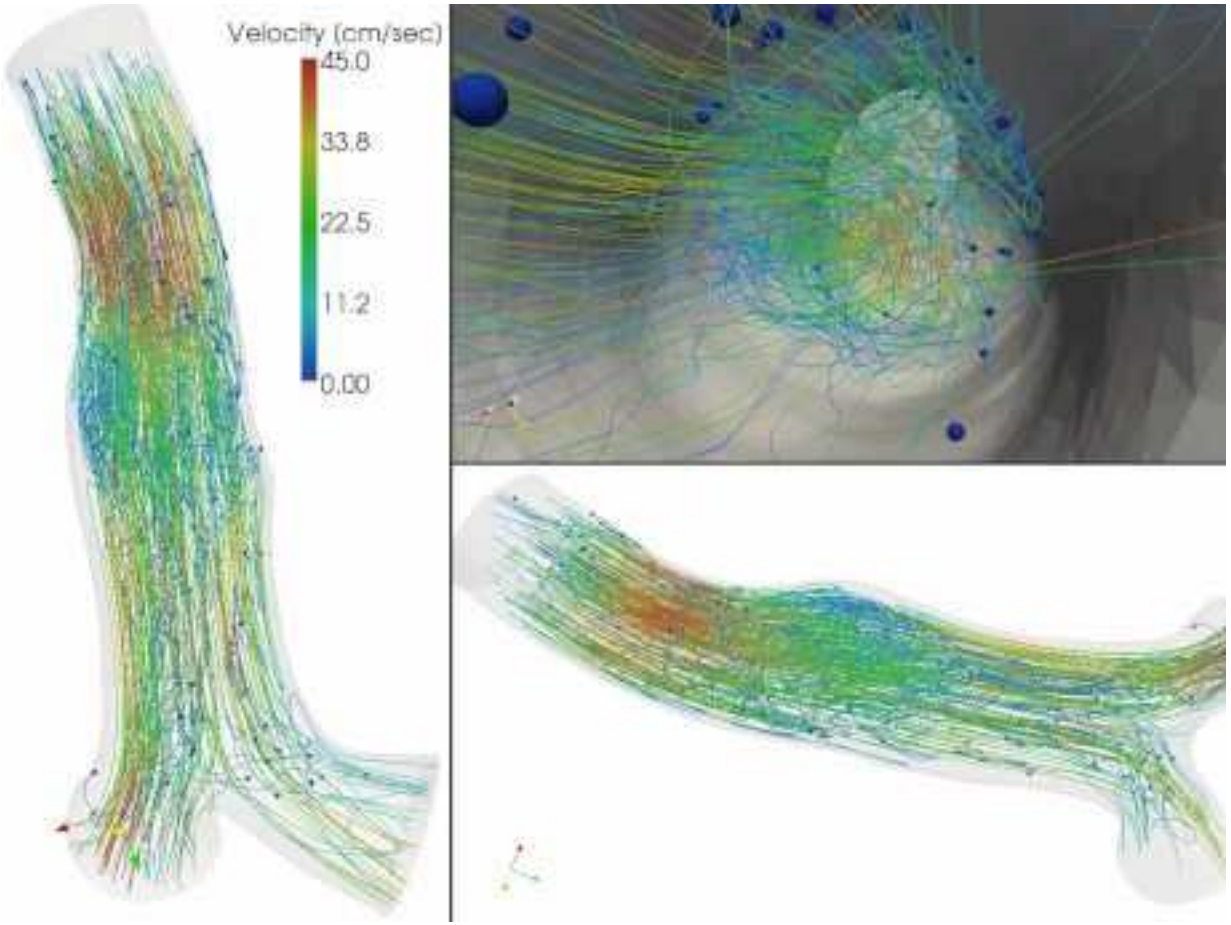
Eulac PerMed 2019

DEXL
DATA EXTREME LAB



Eulac PerMed 2019

DEXL
DATA EXTREME LAB



Eulac PerMed 2019

DEXL
DATA EXTREME LAB

SAVIME System – in-house DBMS

- In-memory
- Multi-dimensional array data model
- Shared Memory Architecture
- Column-store (each variable store in a different dataset)
- Arrays subdivided in subarrays
  - Distributed allocation of subarrays
  - Parallelism inter subarrays and intra elements of the subarray
- Functional Query Language
- Data Ingestion without transformation
- Query Optimization

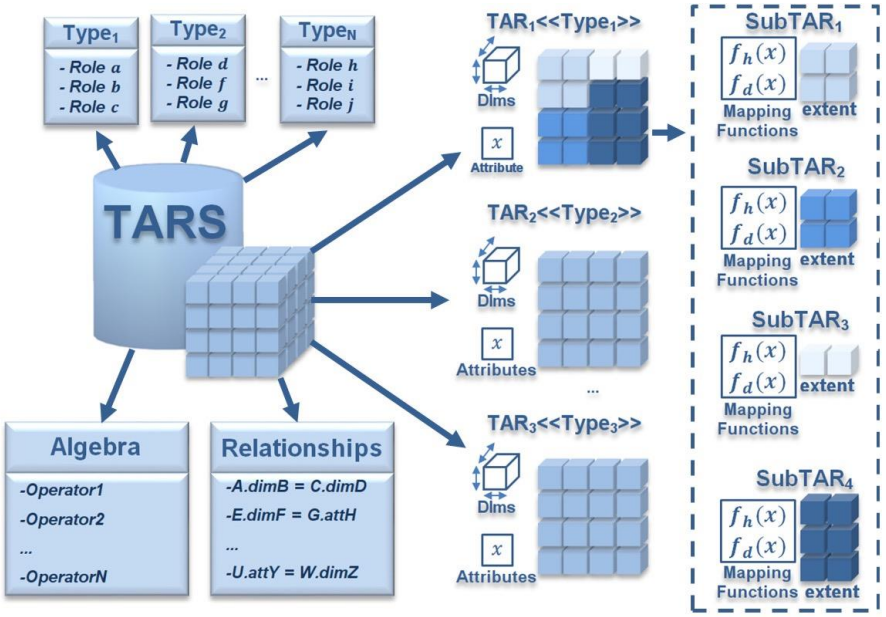
https://github.com/hllustosa/Savime

Hermano Lustosa, Fábio Porto, Pablo Blanco, Patrick Valduriez: Database System Support of Simulation Data. PVLDB 9(13): 1329-1340 (2016)

SAVIME: A Database Management System for Simulation Data Analysis and Visualization, SBBD 2019 (to appear)

DEXL DATA EXTREME LAB

Savime's TARS Data Model



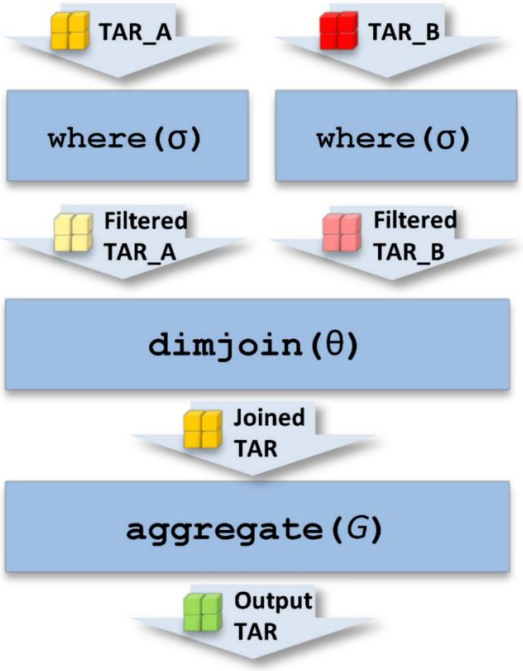
- Typed ARrays Model
- Allows for fast data ingestion
- Regular or Irregular Tiling
- Representation for domain specific data (Simulation Data, Machine Learning Training Data, ...)

Hermano Lustosa, Fábio Porto, Noel Moreno Lemus, Patrick Valduriez: ***TARS: An Array Model with Rich Semantics for Multidimensional Data.*** *ER Forum/Demos 2017: 114-127*

DEXL
DATA EXTREME LAB

Savime's Query Language

```
aggregate(
        dimjoin(
                where(TAR_A,
                predicate),
                where(TAR_B,
                predicate)
        ),
        AVG
);
```



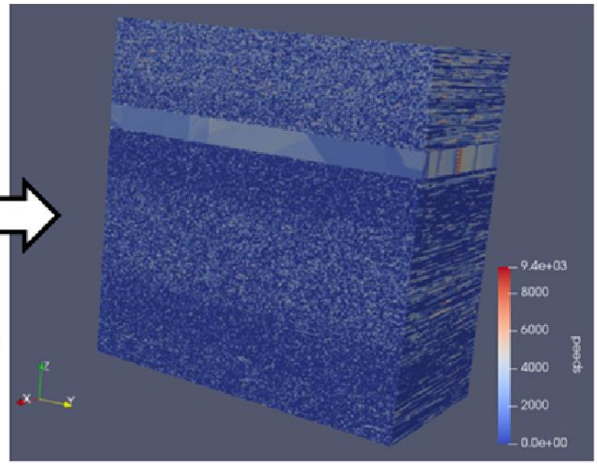
- Functional Query Language with Array Operators
- Array Tiles are pipelined accross array operators

Hermano Lustosa, Fábio Porto, Noel Moreno Lemus, Patrick Valduriez: ***TARS: An Array Model with Rich Semantics for Multidimensional Data.*** *ER Forum/Demos 2017: 114-127*

DEXL DATA EXTREME LAB

Savime's Declarative Visualization

```
catalyze(
  FIELD_DATA_TAR,
  GEOMETRY_TAR,
  TOPOLOGY_TAR,
  'gradient_paraview_script.py'
);
```

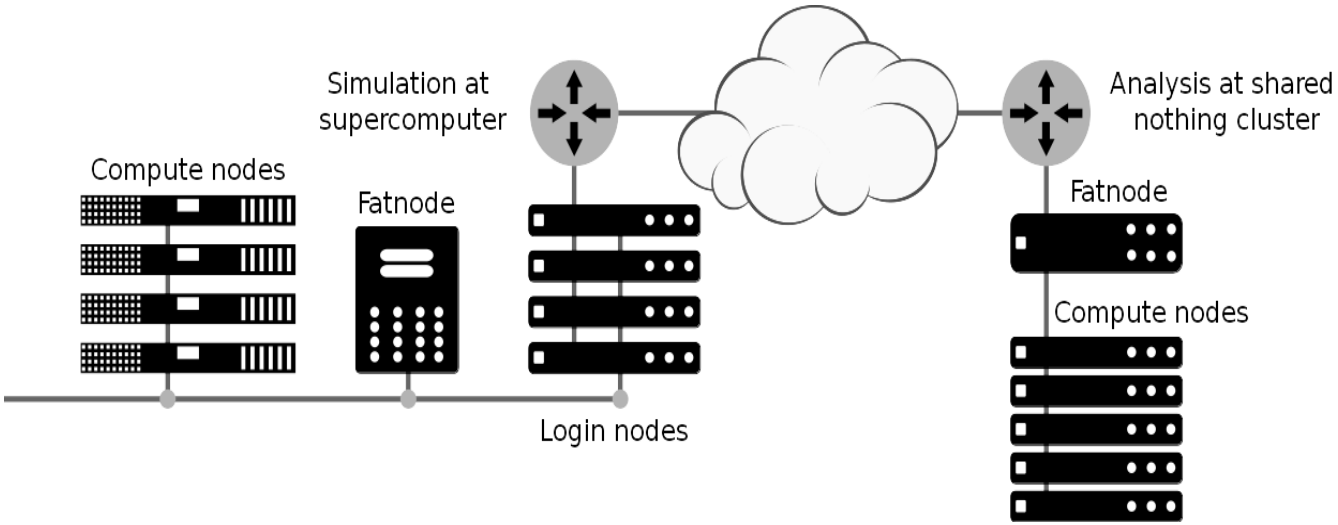


- Special Catalyze Operator outputs query results as a visualization
- Allows Savime integration with Catalyst Library

Hermano Lustosa, Fábio Porto, Noel Moreno Lemus, Patrick Valduriez: ***TARS: An Array Model with Rich Semantics for Multidimensional Data.*** *ER Forum/Demos 2017: 114-127*

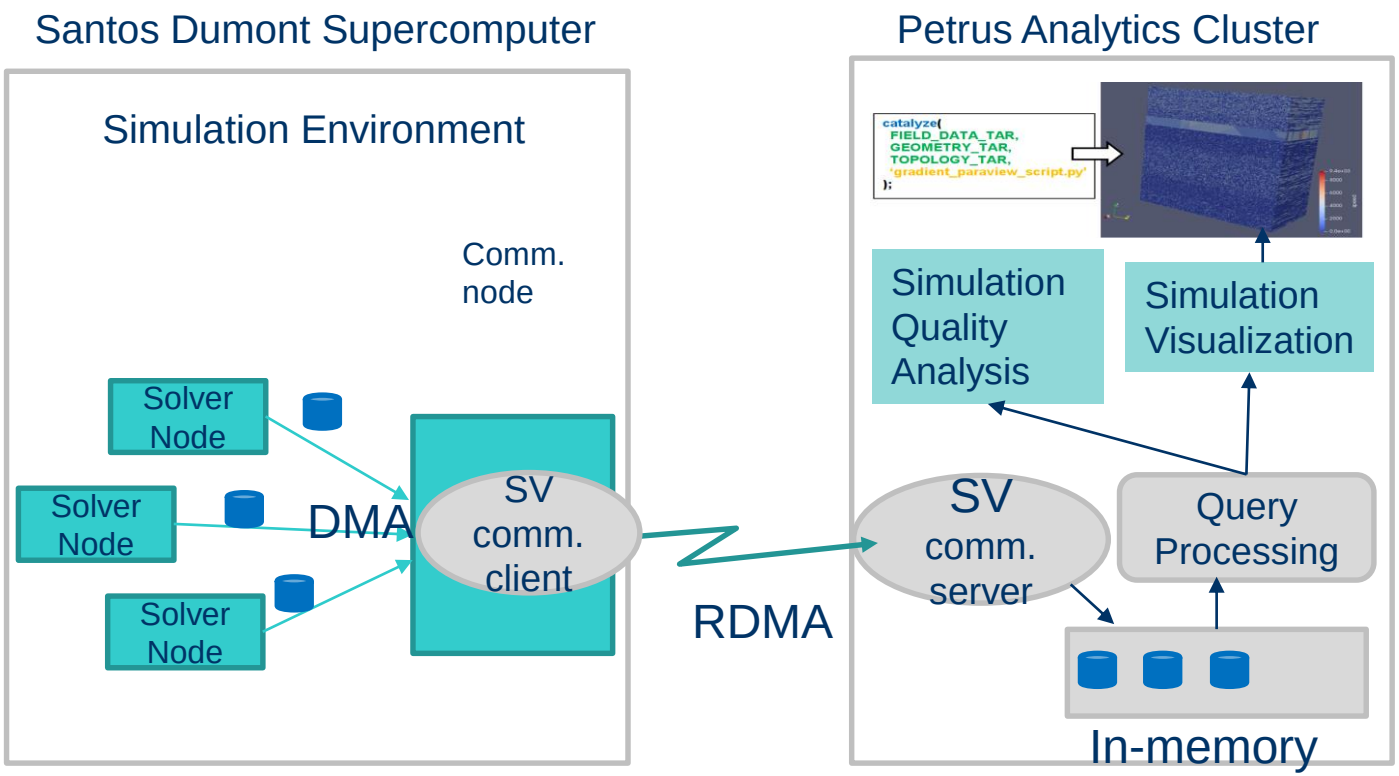
DEXL DATA EXTREME LAB

Integration with the Santos Dumont Supercomputer



DEXL
DATA EXTREME LAB

SAVIME Architecture @LNCC



DEXL DATA EXTREME LAB

Final Comments

- The DEXL Lab has been working on different tracks (methods and tools) to support knowledge structuring and retrieval
- Current Topics of Research
  - Learning with small data
  - Multi-modal learning
  - Physics integrated learning process
  - Automatic selection of spatial models

This is a DEXL Team work

Eulac PerMed 2019

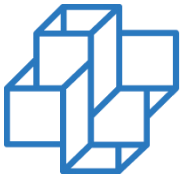
DEXL
DATA EXTREME LAB

Acknowledgements

CNPq
Conselho Nacional de Desenvolvimento Científico e Tecnológico

PETROBRAS

Inria



Obrigado !☺

Fabio Porto
fporto@lncc.br

http://dexl.lncc.br

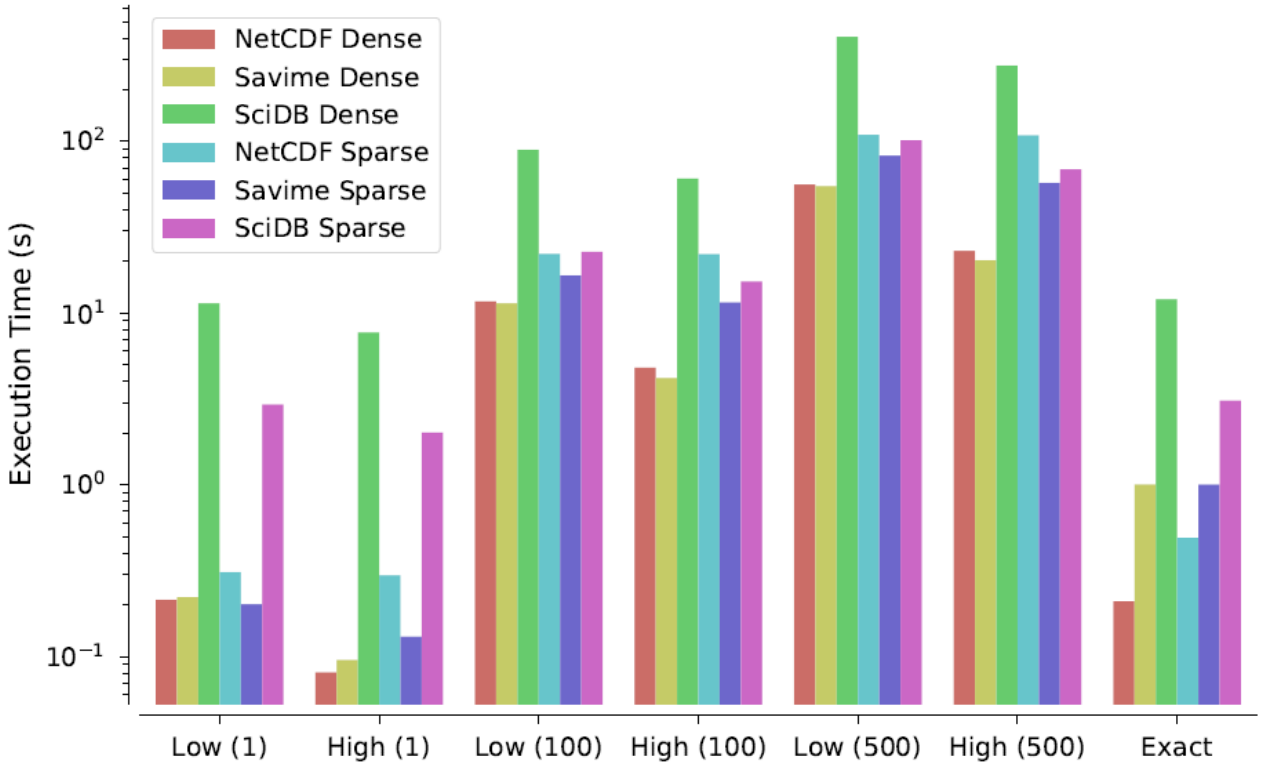
Laboratório Nacional de Computação Científica

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA, INOVAÇÕES E COMUNICAÇÕES

BRASIL GOVERNO FEDERAL

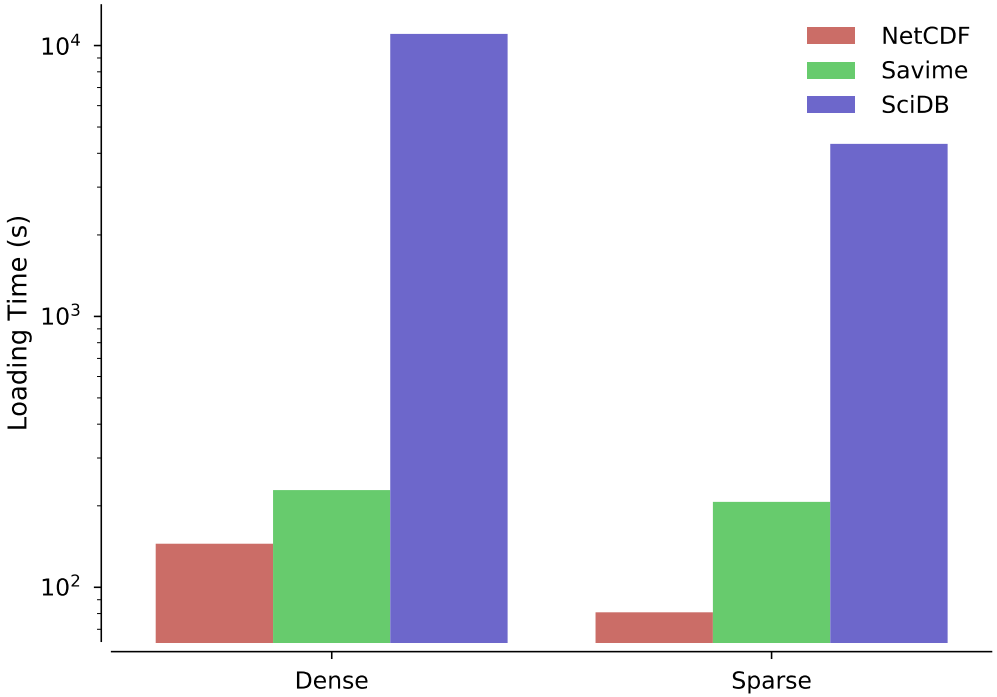
DEXL DATA EXTREME LAB

A Glimpse on Experimental Results



DEXL
DATA EXTREME LAB

Comparison with Loading Time



DEXL
DATA EXTREME LAB