INCT-CID

Fabio Porto
LNCC – MCTIC – DEXL LAB

Semestre 1 – Setembro 2016

Programa de Pós-Graduação em Ciência da Computação

Escola de Informática & Computação
CEFET/RJ

Agenda

- Introdução
- Os Primeiros Passos
- O Turco
- Aplicações

Laboratório Nacional de Computação Científica (LNCC)

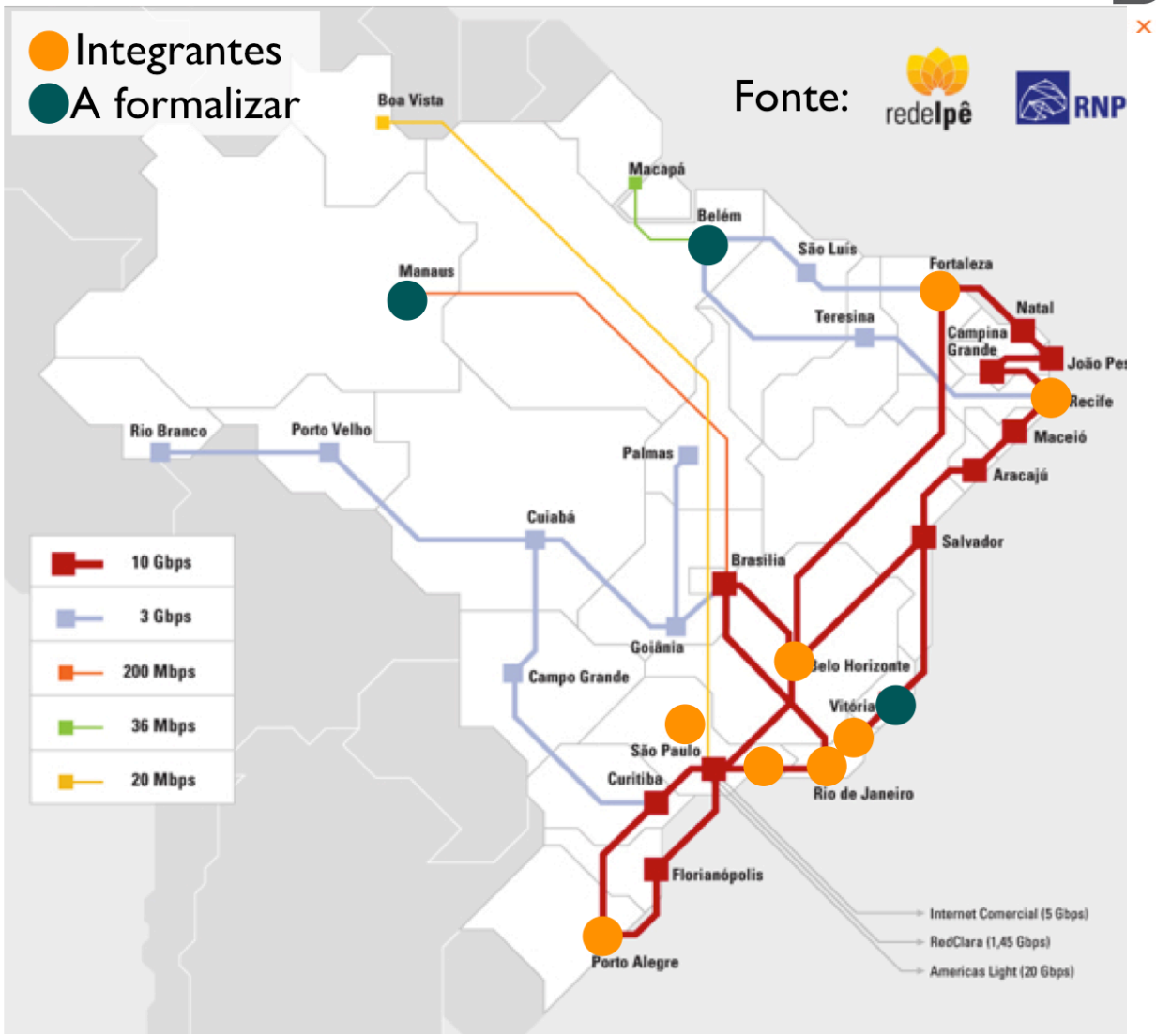
Petropolis, Rio de Janeiro

LNCC - MCTI

- Graduate Course on Computational Modelling
  - CAPES 6
- LABINFO
  - Genomics, proteomics bacteria and micro-organisms
- INCT –MACC
  - Medicine Supported by Scientific Computing
- INCT- CID
- SINAPAD
  - National System of High Performance Performance
- Thematic Laboratory
  - ACIMA (acima.lncc.br)
  - MARTIN (martin.lncc.br)
  - DEXL (dexl.lncc.br)
  - COMCIDIS (comcidis.lncc.br)
  - HEMOLAB (hemolab.lncc.br)
  - LABINFO (labinfo.lncc.br)

SINAPAD – National Center for HPC

- Organized in CENAPADS:
  - Universities
  - Research venters
  - Different architectures:
    - Shared Disks
    - Shared Memory
    - GPUs



Atos-BULL Supercomputer

SANTOS DUMONT CONFIGURATION

Machine Partition	Blades	CPU Processors/Blade	CPU Cores/processor	Total Cores CPU
B710 CPU's	504	2	12	12.096
B715+GPU	396	1	12	4.752
B715+PHI	108	1	12	1.296
B710+B715 CPU's	1008		12	18.144
B710cp+B715cpuPhi	612		12	13.392
MESCLA node			16	250
GPU/PHI	3,7			
	Blades	GPU Processors/Blade	GPU Cores/processor	Total Cores GPU or PHI
NVIDIA K40	396	1	2880	1.140.480
INTEL 7120X	108	1	61	6.588

Grupo de Pesquisa em Ciência de Dados

- Artur Ziviani
- Eduardo Ogasawara (CEFET-RJ)
- Fabio Porto
- Kary Ocana
- Luiz Gadelha
- Andre Salles

Ciência de Dados – Uma nova Disciplina

- Mudança de Visão Histórica para Predição do Futuro
  - Predições razas e profundas
- Inferências baseadas em dados
- Modelos Baseadas em dados
- Foco
  - Grande volume
  - Treinamento
  - Tratamento
  - Interpretação Humana

INCT - Ciência de Dados (CID)

- Objetivo
  - Estruturar a nova área
    - Formação de Recursos Humanos
    - Pesquisa e Desenvolvimento
    - Transferência de Tecnologia para Indústria e Sociedade
  - Como?
    - Através da Integração de disciplinas com foco em dados
      - Bancos de Dados
      - Análise de Dados
      - Modelagem de Dados
    - Metodologias
      - Dirigida à hipóteses
      - Análise estatística
    - Tecnologias
      - Frameworks Big Data: Spark, Giraph, Tensor Flow, ….
      - Linguagens de programacão: R, python, scala
      - Sistemas: NoSQL, Greenplum, ...
      - Machine Learning: scikit-learn, Apache Mahout,..

INCT - CID

- Parcerias
  - FIOCRUZ
  - Observatório Nacional
  - Comitê Olimpico Brasileiro
  - Prefeituras: Curitiba, Rio de Janeiro
- Empresas
  - EMC - Centro de Pesquisa
  - IBM Research
  - Philips Research
- Conselho Internacional
  - NYU
  - INRIA
  - UCSB
  - Boston University – Data Science Institute

Uma breve história da Ciência de Dados

- *R.A. Fisher,* 1935, *The design of Experiments*

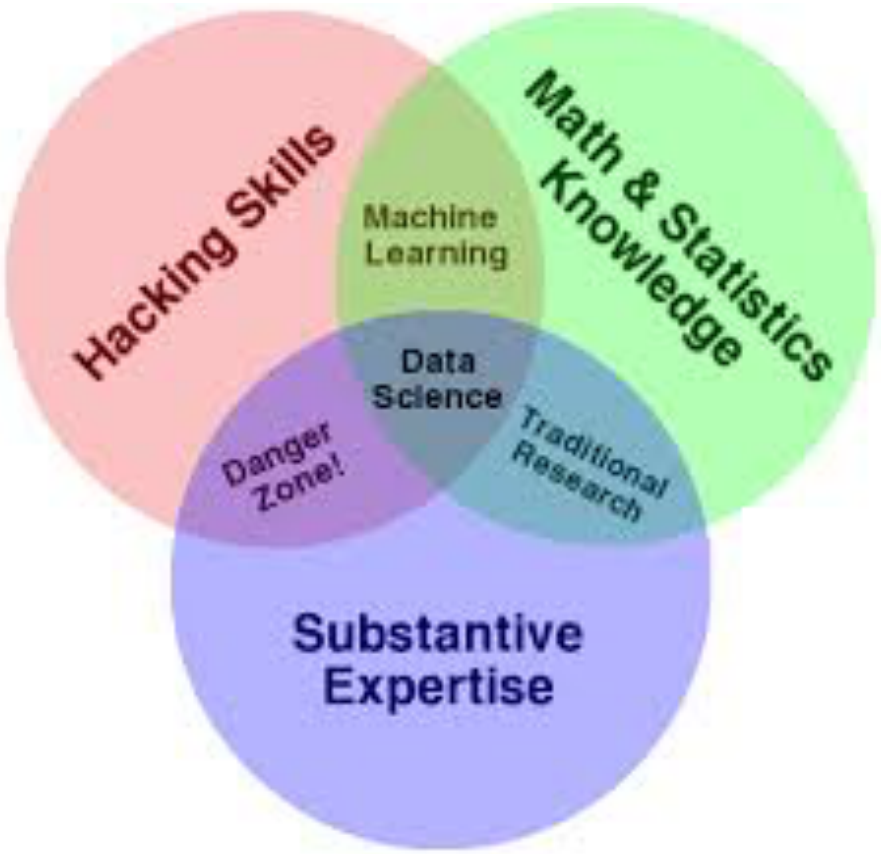
  Correlation does not imply causation

- *Hans P. Luhn, 1958, Business Intelligent Systems, IBM*
  - *Automatic method to provide current awareness services to scientists and engineers*
- *J.W. Tukey, 1977, Exploratory Data Analysis*
  - *.. For seeing what the data can tell us beyond the formal modeling and hypothesis testing .. – inspirou o desenvolvimento do pacote estatítico S*
- *H. Dresner,1989, Business Intelligence*
  - *The database view on data analytics*

Uma breve história da Ciência de Dados

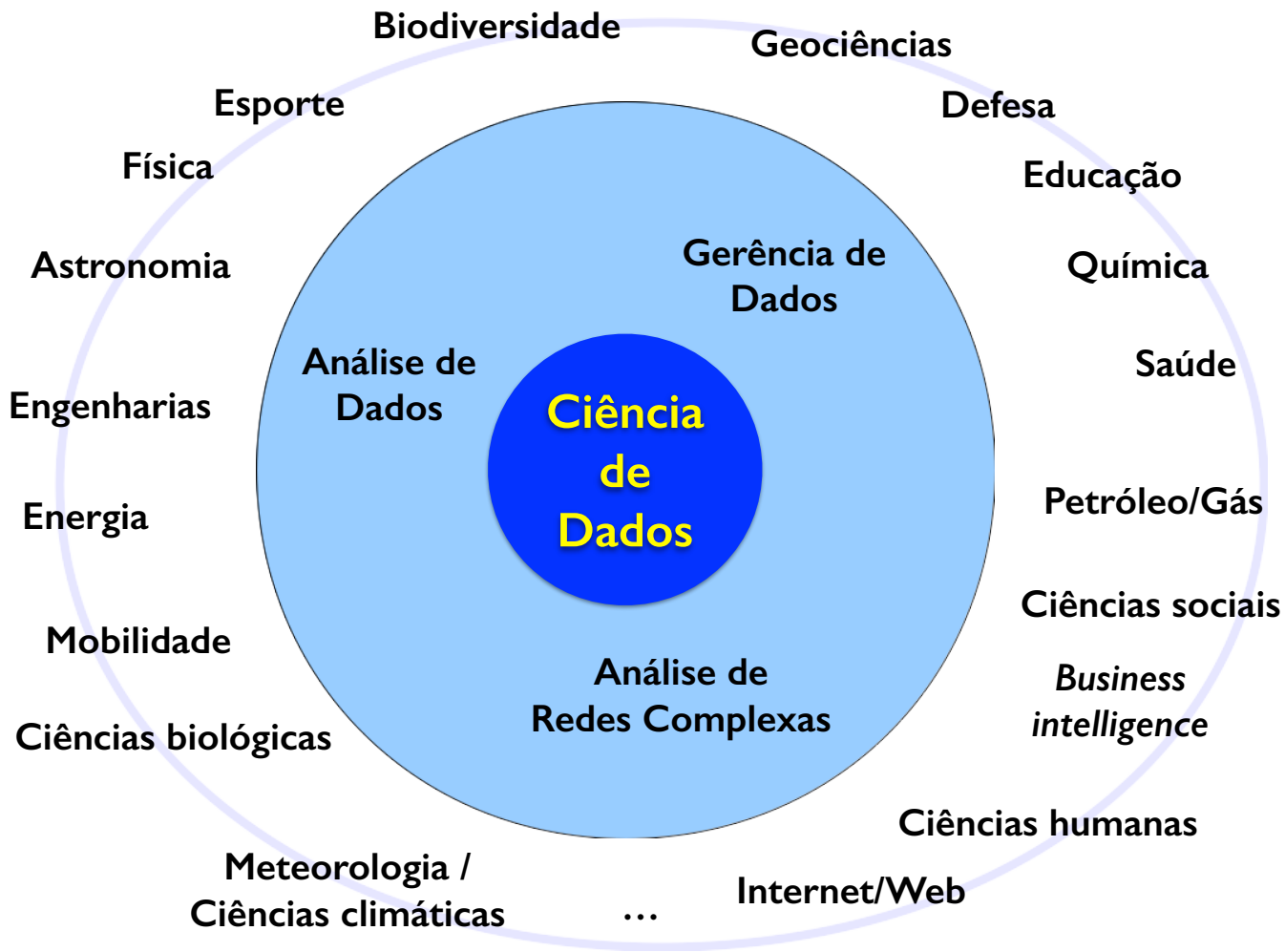
- *T. Michel, 1977, The Machine Learning book*
- *Google, 1996, Prototype Search Engine*
- *Jim Gray et. al, 2007, The Fourth Paradigm*
  - *Experimental data science*
- *A. Halevy, P. Norvig, 2009, The Unreasonable Effectiveness of Data*
  - *From deep models to harnessing of data volumes*
- *Exponential growth in data volume, 2010, The data deluge*

Data Science - Habilidades



Fonte: http://drewconway.com/

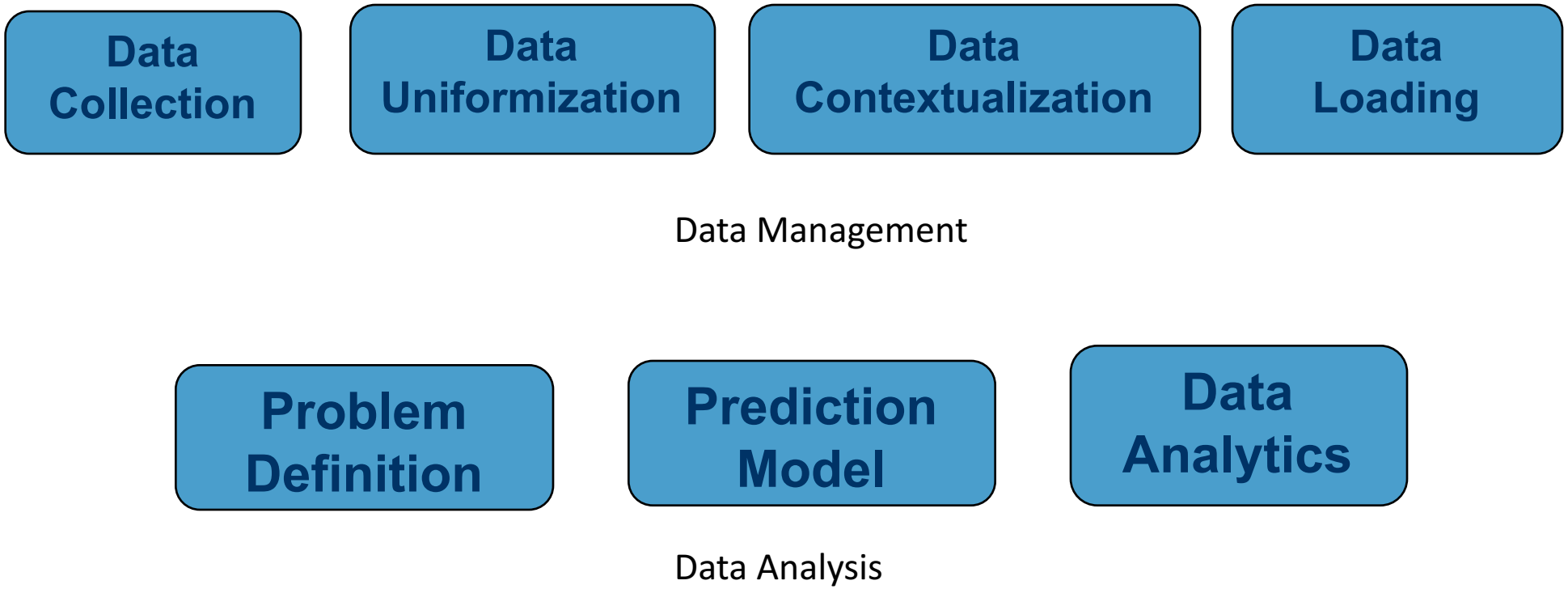
Organização do INCT-CID



CS + X

- Ciência da Computação está na base do processo científico em todas disciplinas
  - CS + astronomia
  - CS + biologia
  - CS + meteorologia
  - CS + esporte
  - CS + sociologia
- you name it !!
- Fundamental na ciência de dados
- *The Fourth Paradigma: Data Intensive Scientific Discovery, Jim Gray et al.*

Processo de CID



De Dados Crus à Interpretação

DEXL LAB
EXTREME DATA LAB

cleaning

Integration

Transformation

Analysis

Raw

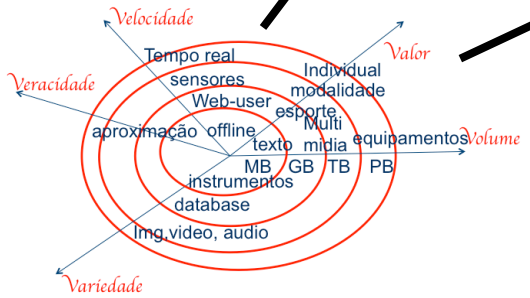
Tier:0

Intermediary

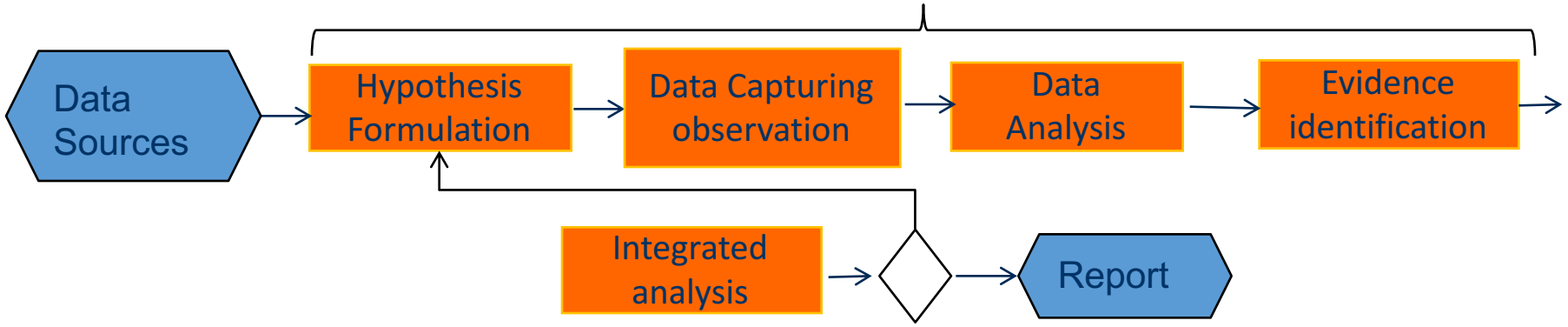
Tier:1

Tailored

Tier:2

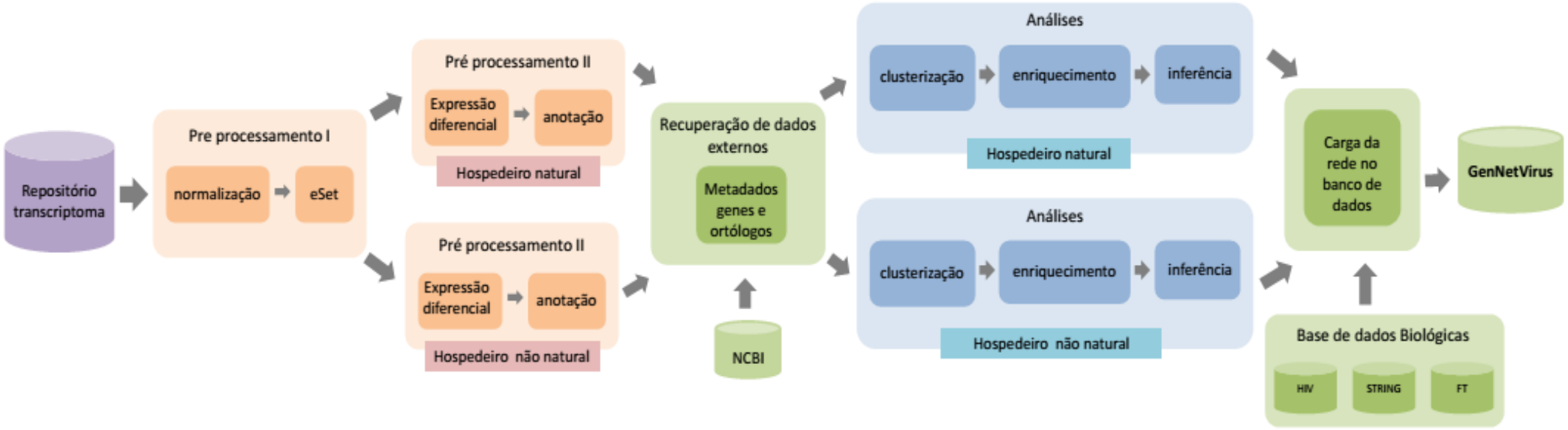


Ciência de Dados – Processo

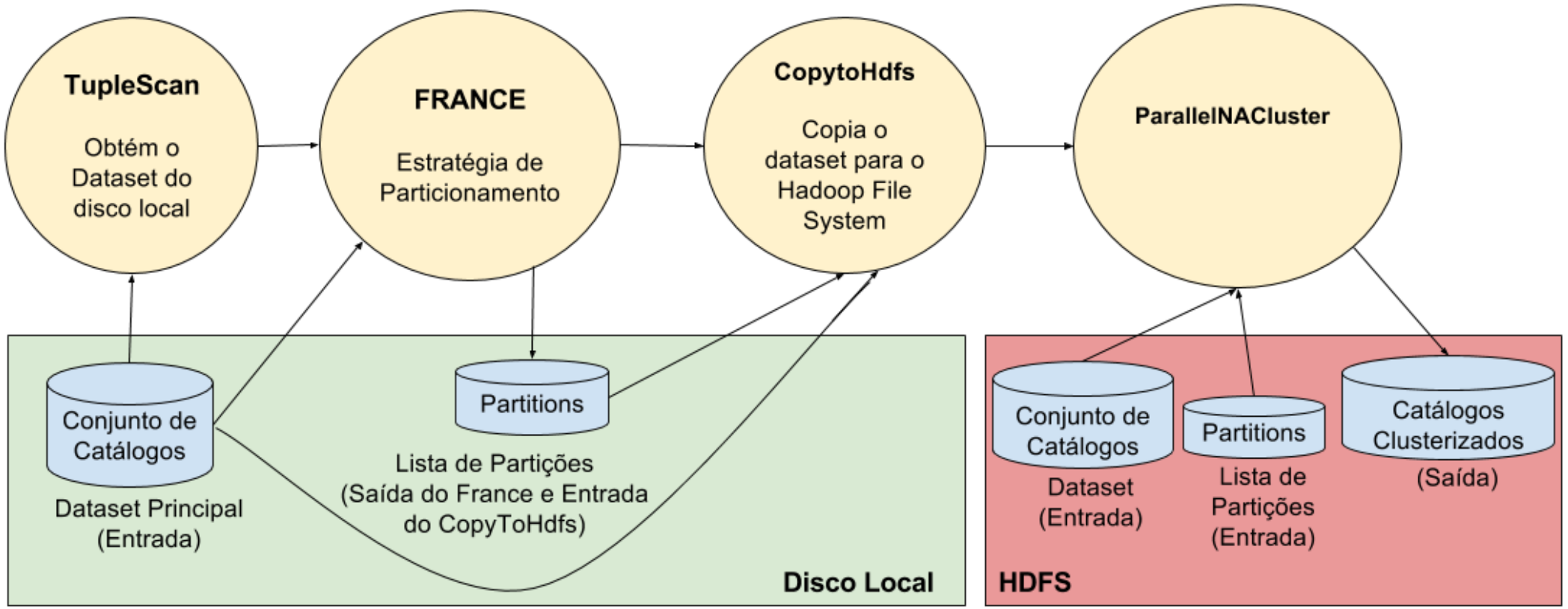


Um dataflow para carga de dados consolidado – GenNetVirus (R. L. Costa)

DEXL LAB
EXTREME DATA LAB

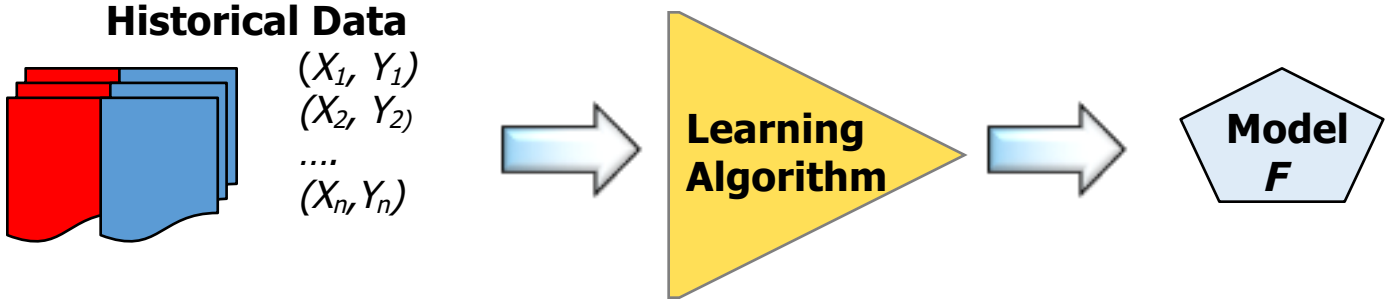


NaCluster – V. Pires

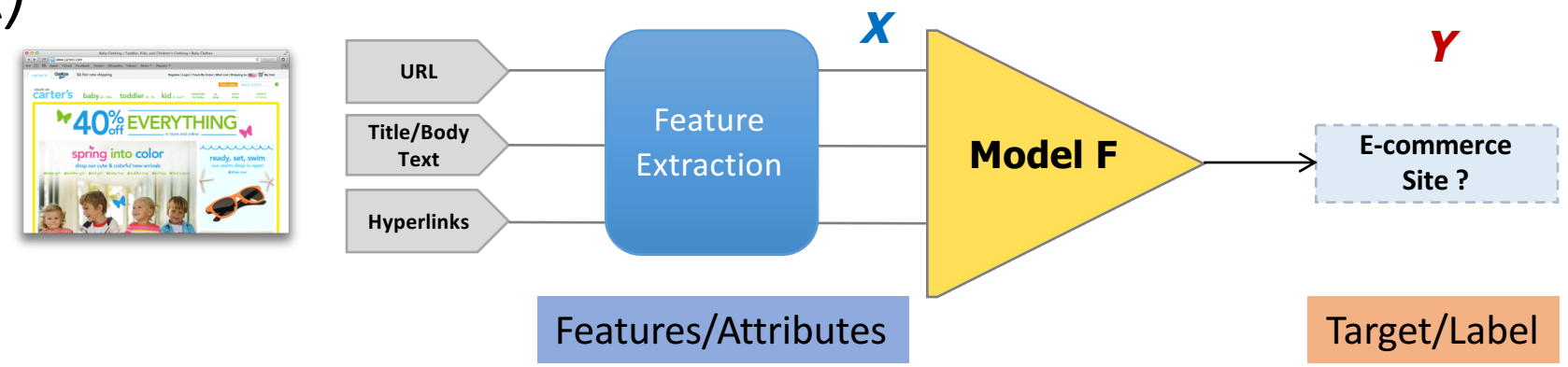


Supervised Learning

- **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)

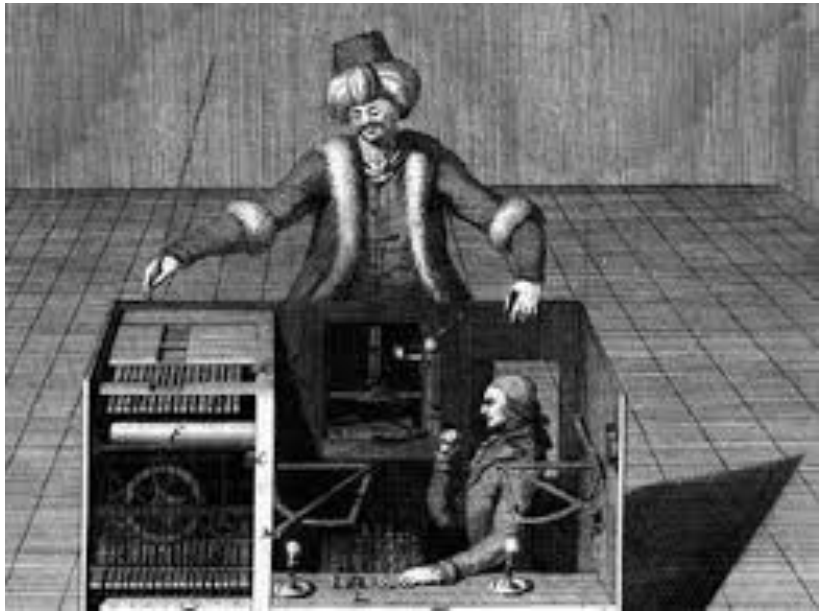


- **Prediction:** Given a new sample X with unknown Y, predict Y using $F(X)$

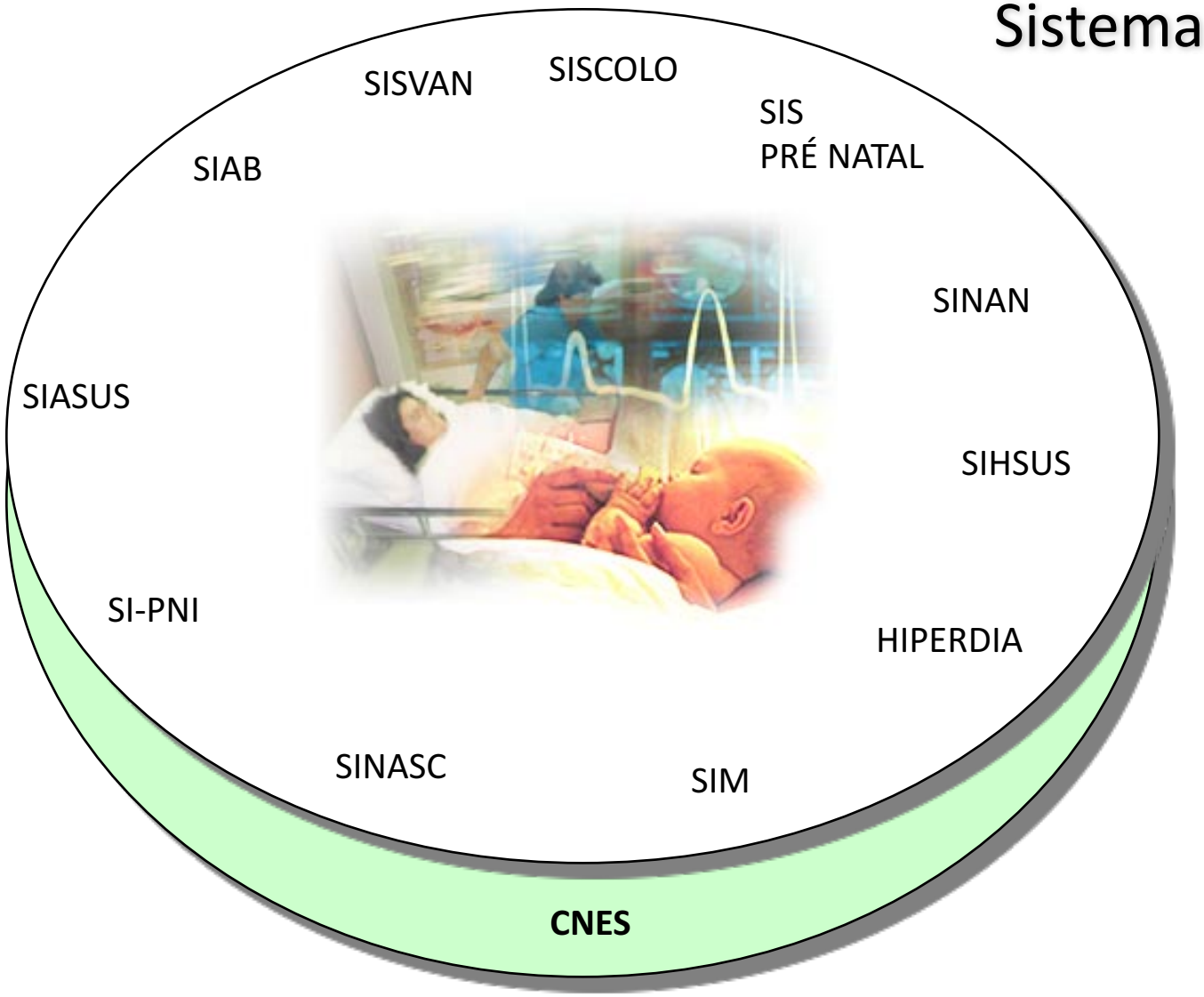


Human Data Analytics

DEXL LAB
EXTREME DATA LAB

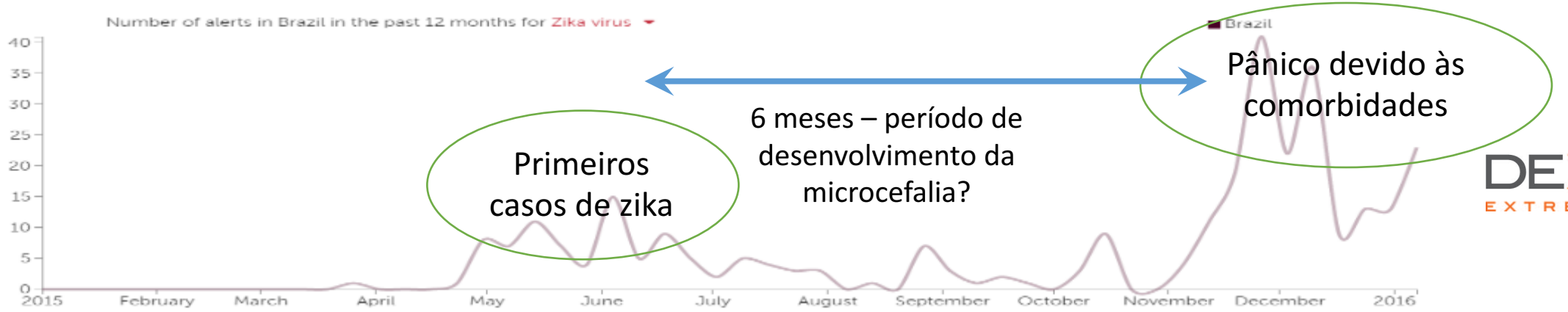


Sistemas de Informação de Saúde



Registram ‘eventos’, que ocorrem em pessoas, não os processos de adoecimento e morte

Linkagem de dados, possível mas difícil

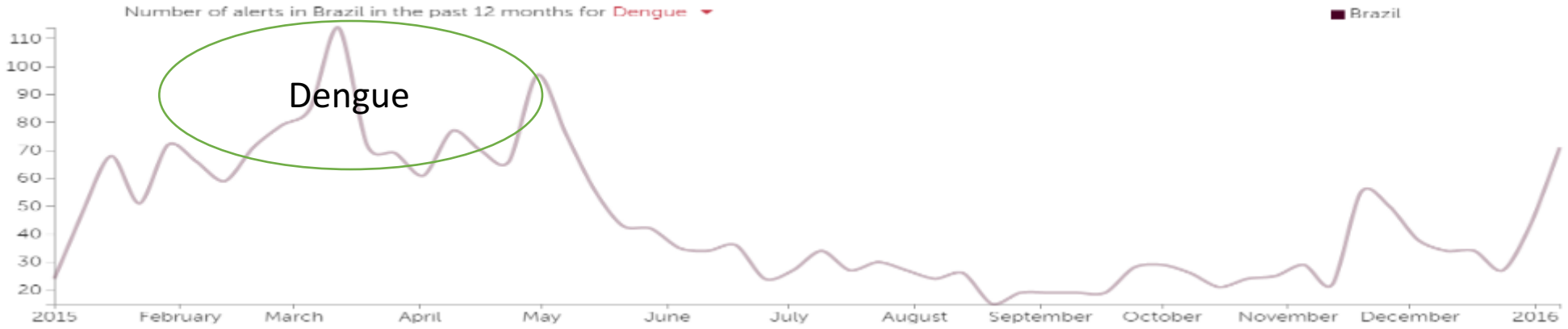


Primeiros casos de zika

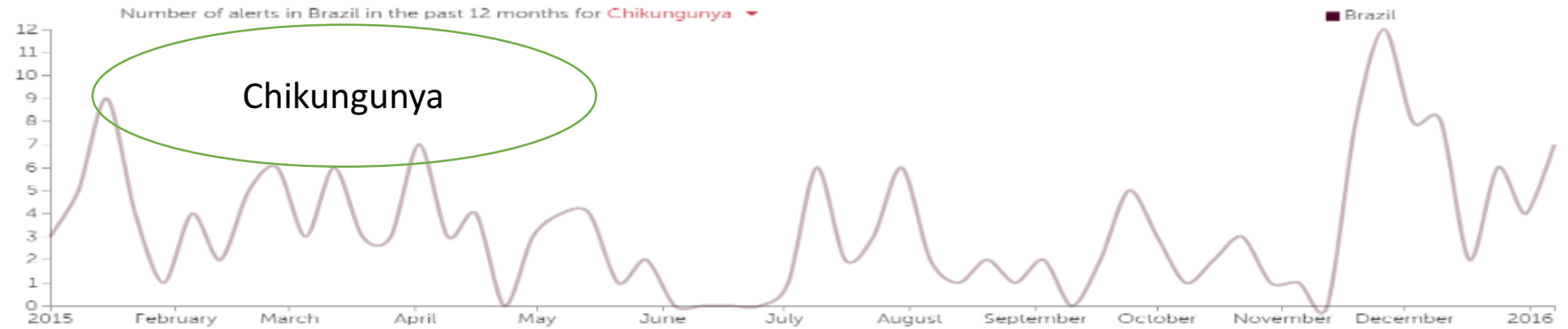
6 meses – período de desenvolvimento da microcefalia?

Pânico devido às comorbidades

DEXL LAB
EXTREME DATA LAB



Dengue

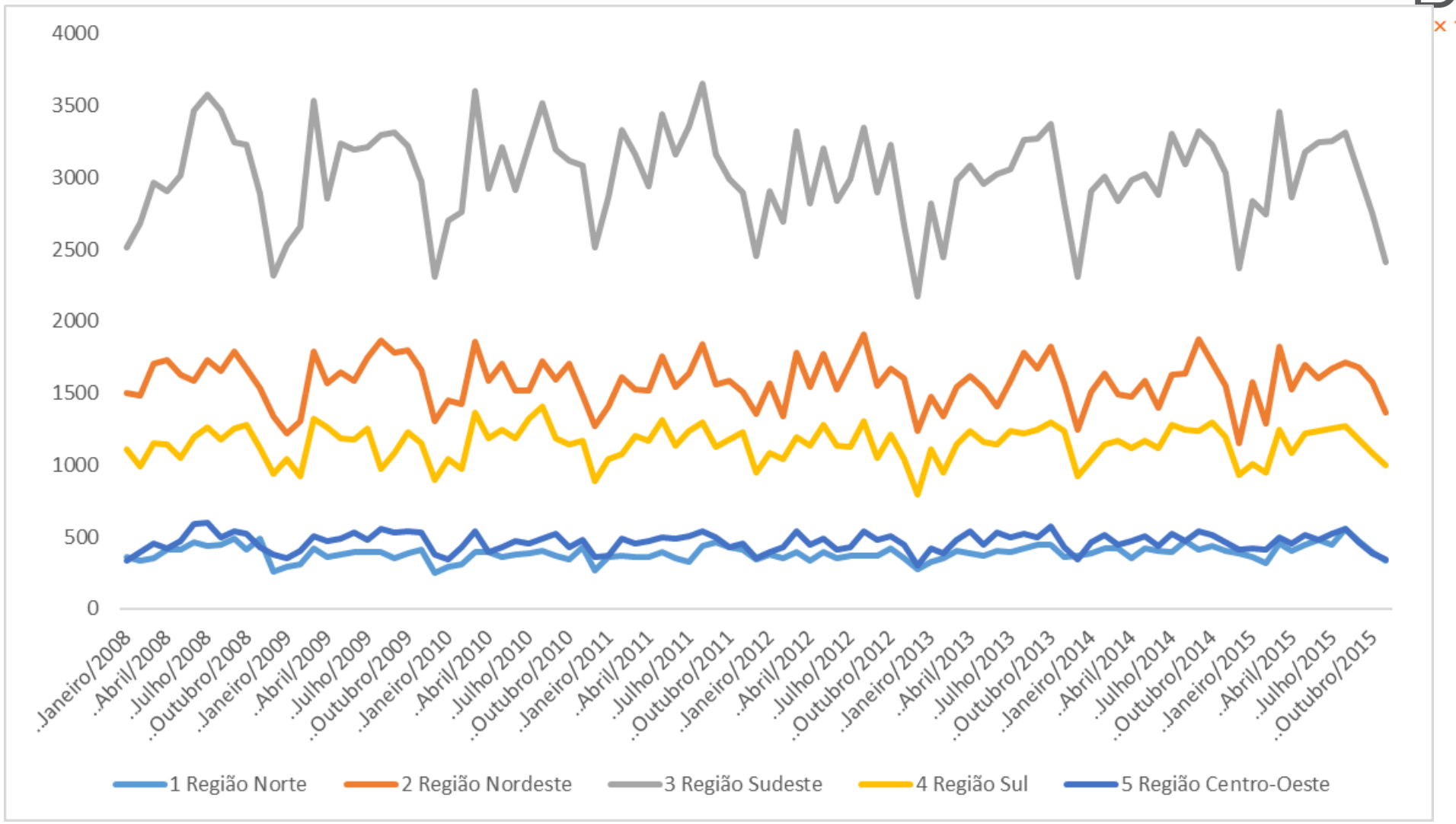


Chikungunya

Lista de doenças possivelmente associadas à infecção pelo vírus zika

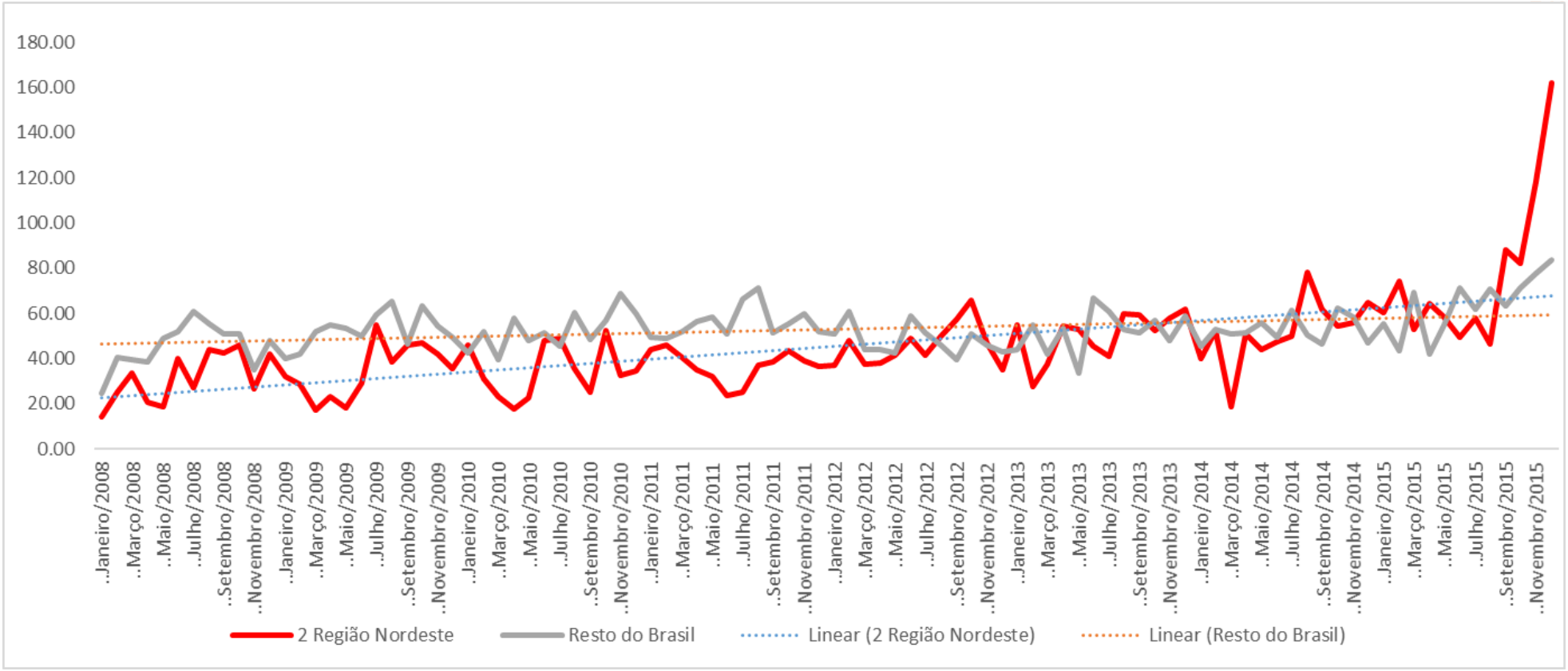
	Doença ou condição	Evento	Sistema	CID-10
INFECÇÕES VIRAIS	Dengue (dengue clássico)	Notificação	SINAN	A90
	Febre hemorrágica devida ao vírus da dengue	Notificação, óbito	SINAN, SIM	A91
	Febre de Chikungunya	Notificação	SINAN	A92.0
	Doença pelo Zika vírus	Notificação	SINAN	U06*
	Doença pelo Zika vírus, não especificada		SINAN	U06.9*
MANIFESTAÇÕES NEUROLÓGICAS ASSOCIADAS A INFECÇÕES VIRAIS (DENGUE, CHIKUNGUNYA E ZIKA)	Encefalite por vírus transmitidos por mosquitos	Internação hospitalar, óbito	SIH	A83
	Outras encefalites por vírus transmitidas por mosquitos		SIH	A83.8
	Encefalite não especificada por vírus transmitida por mosquitos		SIH	A83.9
	Outras encefalites virais, não classificadas em outra parte	Internação hospitalar	SIH	A85
	Encefalite viral não especificada	Internação hospitalar	SIH	A86
	Meningite viral	Internação hospitalar	SINAN	A87
	Outras meningites virais			A87.8
	Meningite viral não especificada			A87.9
	Meningite em outras doenças infecciosas e parasitárias classificadas em outra parte	Internação hospitalar, óbito	SIH, SIM	G02
	Meningite em doenças virais classificadas em outra parte		SIH, SIM	G02.0
	Meningite devida a outras causas e a causas não especificadas	Internação hospitalar, óbito	SIH, SIM	G03
	Meningite devida a outras causas especificadas		SIH, SIM	G03.8
	Meningite não especificada		SIH, SIM	G03.9
	Encefalite, mielite e encefalomielite	Internação hospitalar, óbito	SIH, SIM	G04
	Encefalite aguda disseminada (ADEM)		SIH, SIM	G04.0
	Outras encefalites, mielites e encefalomielites		SIH, SIM	G04.8
	Encefalite, mielite e encefalomielite não especificada		SIH, SIM	G04.9
	Encefalite, mielite e encefalomielite em doenças classificadas em outra parte	Internação hospitalar, óbito	SIH, SIM	G05
	Encefalite, mielite e encefalomielite em doenças virais classificadas em outra parte		SIH, SIM	G05.1
	Síndrome de Guillain-Barré	Internação hospitalar, óbito	SIH, SIM	G61.0
	Hemiplegia	Internação hospitalar	SIH	G81
	Hemiplegia flácida		SIH	G81.0
	Paraplegia e tetraplegia	Internação hospitalar	SIH	G82
	Paraplegia flácida		SIH	G82.0
	Paraplegia não especificada		SIH	G82.2
	Tetraplegia flácida		SIH	G82.3
	Síndrome paralítica não especificada	Internação hospitalar	SIH	G83.9
TIPOS E CAUSAS DE ABORTO ASSOCIADOS À INFECÇÃO PELO VÍRUS ZIKA	Gravidez ectópica	Internação hospitalar, óbito fetal ou da gestante	SIH, SIM-feto	O00
	Mola hidatiforme	Internação hospitalar, óbito fetal ou da gestante	SIH, SIM-feto	O01
	Outros produtos anormais da concepção	Internação hospitalar, óbito fetal ou da gestante	SIH, SIM-feto	O02
	Aborto espontâneo	Internação hospitalar, óbito fetal ou da gestante	SIH, SIM-feto	O03
	Aborto por razões médicas e legais	Internação hospitalar óbito fetal ou da gestante	SIH, SIM-feto	O04
	Outros tipos de aborto	Internação hospitalar óbito fetal ou da gestante	SIH, SIM-feto	O05
	Aborto não especificado	Internação hospitalar óbito fetal ou da gestante	SIH, SIM-feto	O06
	Falha de tentativa de aborto	Internação hospitalar óbito fetal ou da gestante	SIH, SIM-feto	O07
	Complicações consequentes a aborto e gravidez ectópica ou molar	Internação hospitalar óbito fetal ou da gestante	SIH, SIM-feto	O08
TIPOS DE MALFORMAÇÕES CONGÊNITAS ASSOCIADOS À INFECÇÃO PELO VÍRUS ZIKA	Anencefalia e malformações similares	Nascimento, óbito fetal, internação do neonato	SINASC, SIM-neonato, SIH	Q00
	Microcefalia	Nascimento, óbito fetal, internação do neonato	SINASC, SIM-neonato, SIH	Q02
	Hidrocefalia congênita	Nascimento, óbito fetal, internação do neonato	SINASC, SIM-neonato, SIH	Q03
	Outra hidrocefalia congênita		SINASC, SIM-neonato, SIH	Q03.8
	Hidrocefalia congênita não especificada		SINASC, SIM-neonato, SIH	Q03.9
	Outras malformações congênitas do cérebro	Nascimento, óbito fetal, internação do neonato	SINASC, SIM-neonato, SIH	Q04
	Outras deformidades por redução do encéfalo (incluindo hidranencefalia)		SINASC, SIM-neonato, SIH	Q04.3
	Malformação congênita não especificada do encéfalo		SINASC, SIM-neonato, SIH	Q04.9
	Malformações congênitas das pálpebras, do aparelho lacrimal e da órbita	Nascimento	SINASC	Q10
	Anoftalmia, microftalmia e macroftalmia	Nascimento	SINASC	Q11
	Malformações congênitas do cristalino	Nascimento	SINASC	Q12
	Malformações congênitas da câmara anterior do olho	Nascimento	SINASC	Q13
	Malformações congênitas da câmara posterior do olho	Nascimento	SINASC	Q14
	Outras malformações congênitas do olho	Nascimento	SINASC	Q15
	Malformações congênitas do ouvido causando comprometimento da audição	Nascimento	SINASC	Q16
	Outras malformações congênitas da orelha	Nascimento	SINASC	Q17

Evolução das internações por total de malformações em menores de um ano

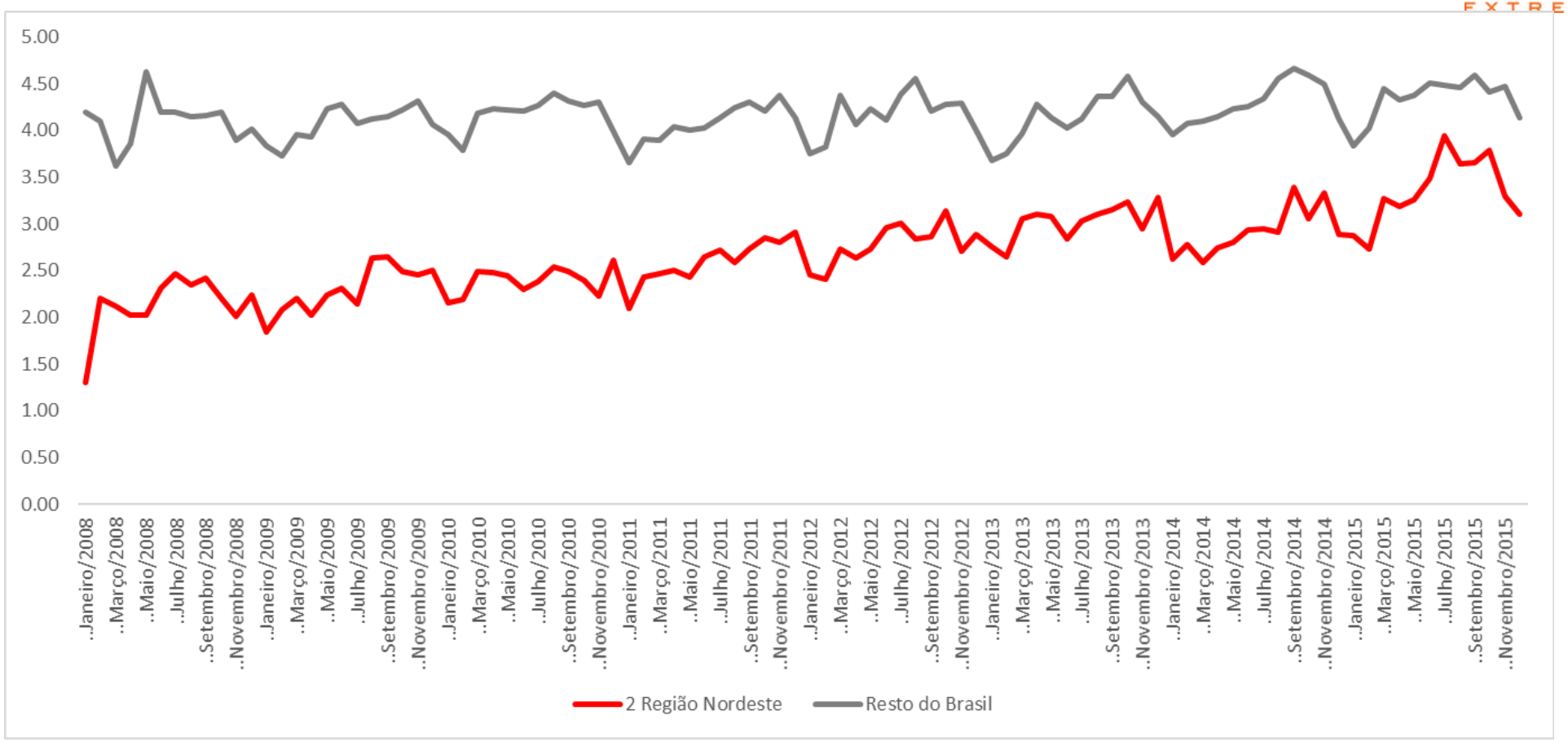


Evolução das internações por malformações do SNC em menores de um ano

DEXL LAB
TREME DATA LAB

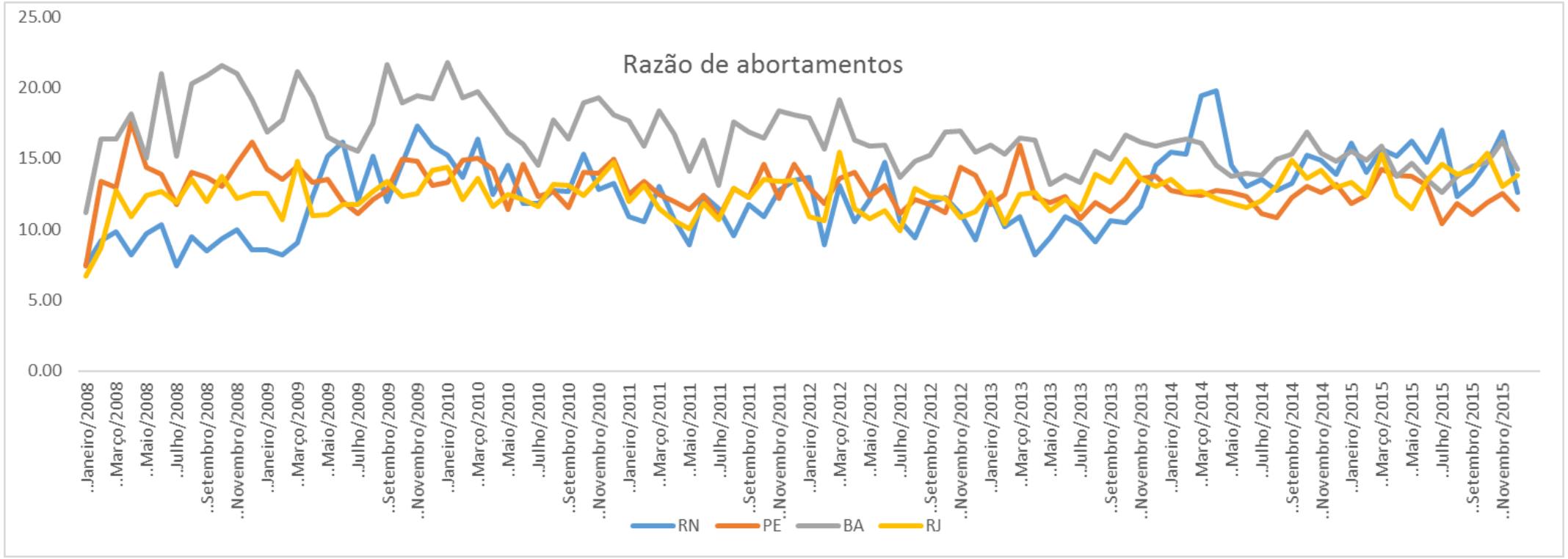


Evolução das internações por “outras doenças do SNC”

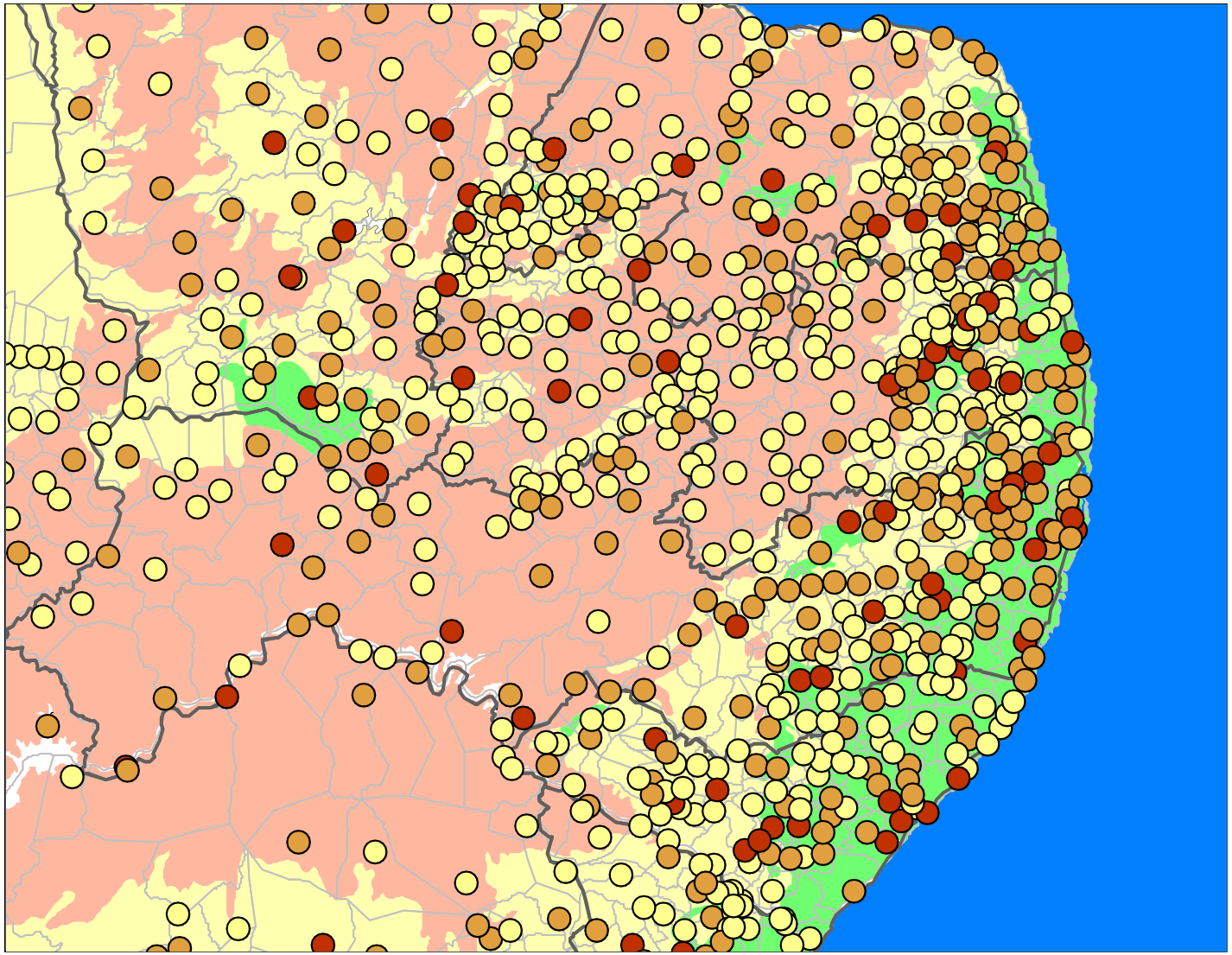


Evolução das internações por complicações relacionadas ao aborto

DEXL LAB
EXTREME DATA LAB



Aumento das internações por neuropatias

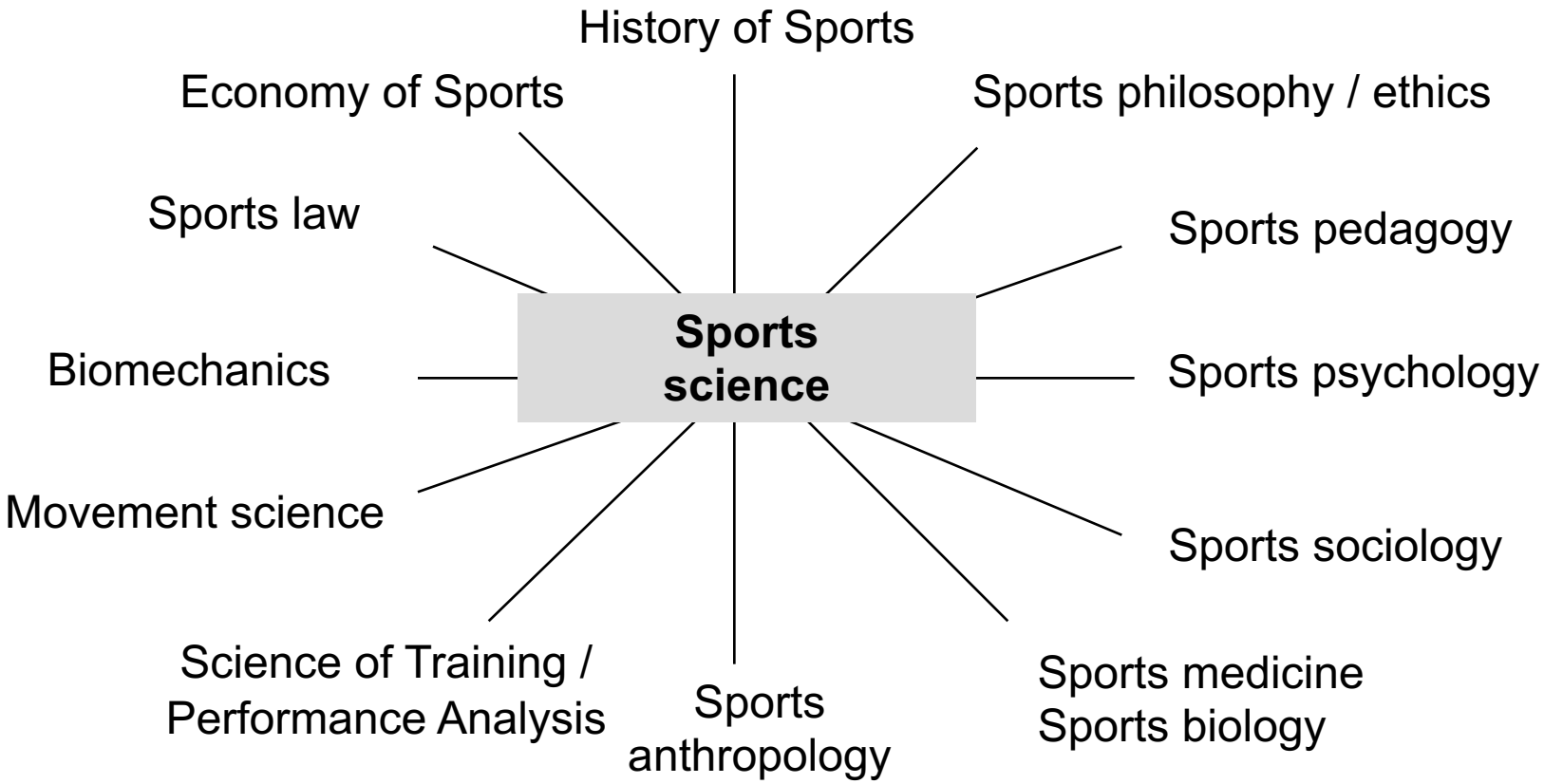


Melhorar o indicador
Usar técnicas de detecção de cluster espacial

DEXL LAB
EXTREME DATA LAB

Área Exemplo: Esporte de Alto Rendimento

Áreas da Ciência do Esporte

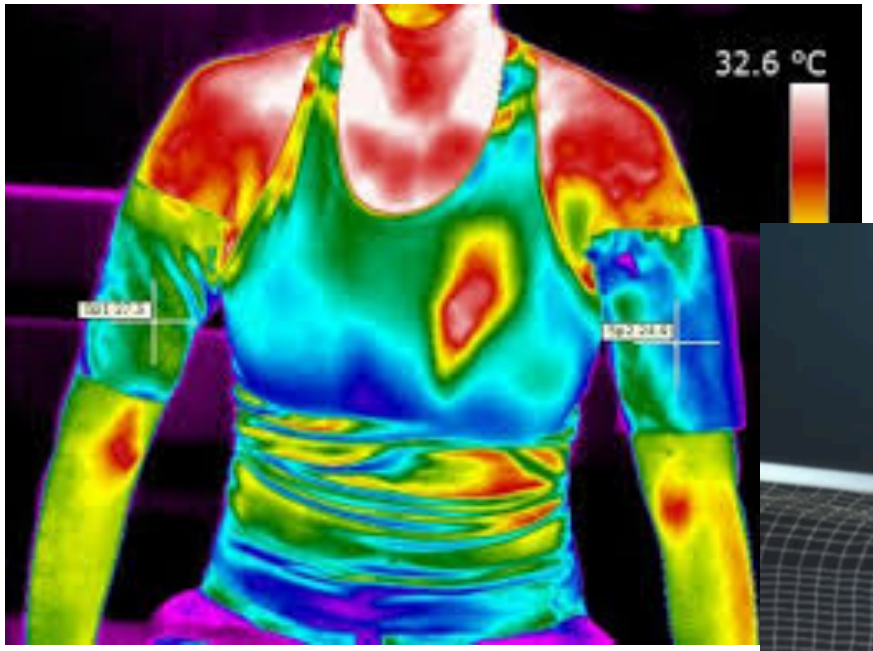


Sportomics [Cameron, Bassini 2015]

- Adoção da Avaliação integrada de Atletas a partir de diferentes disciplinas
- Características individuais - > fazem a diferença
  - características raras são mais valiosas !!!
- Processo científico baseado em dados

Computação: atleta e sua modalidade

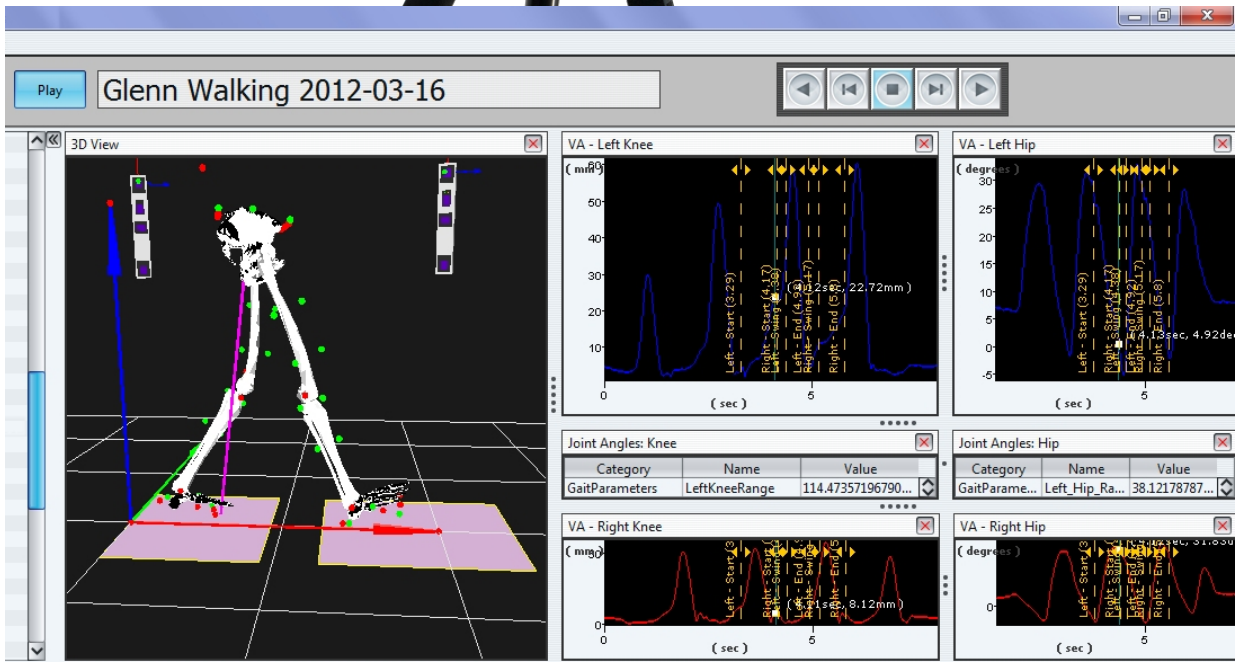
DEXL LAB
EXTREME DATA LAB



Captura de Dados

codamotion

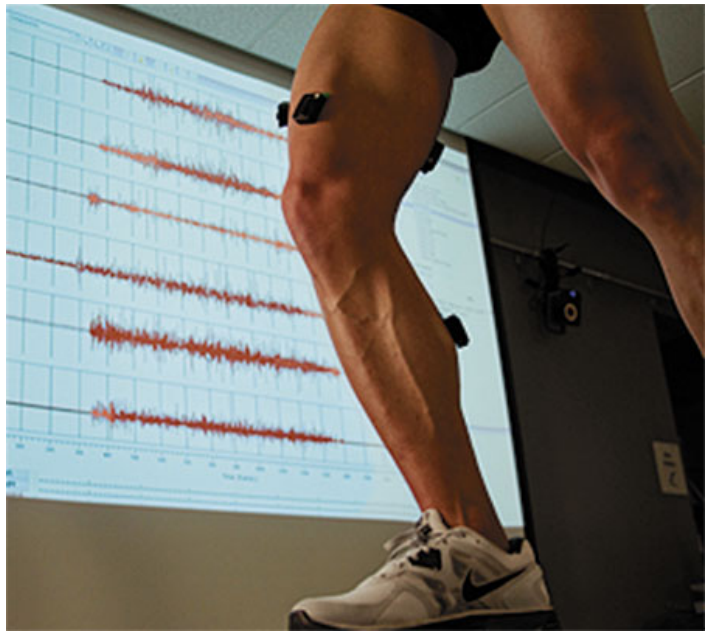
EXL LAB
REME DATA LAB



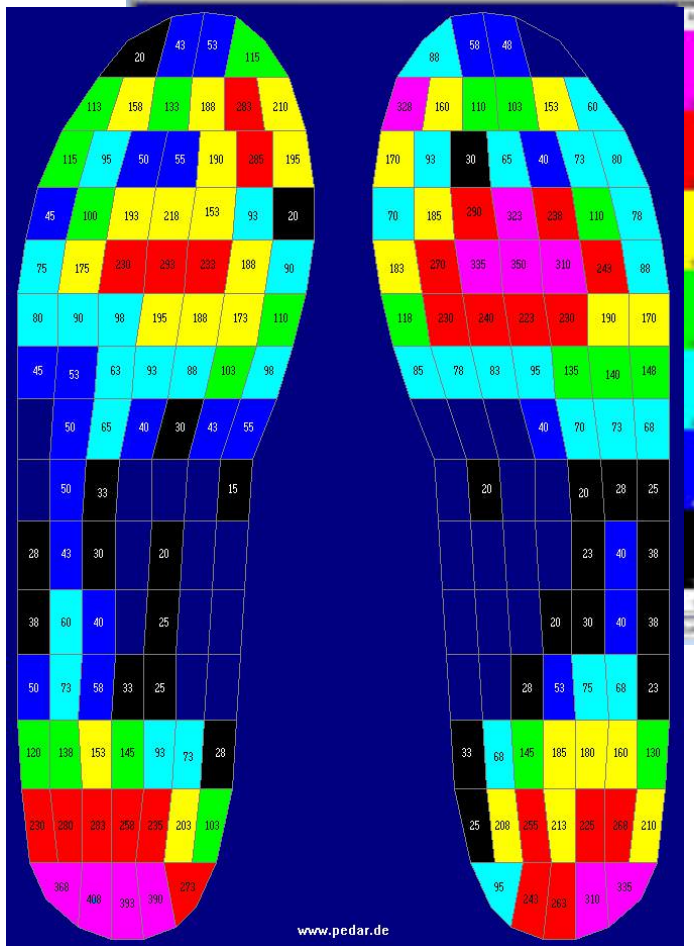
DELSYS®

WEARABLE SENSORS
FOR MOVEMENT SCIENCES

DEXL LAB
EXTREME DATA LAB



novel.de



XL LAB
ME DATA LAB

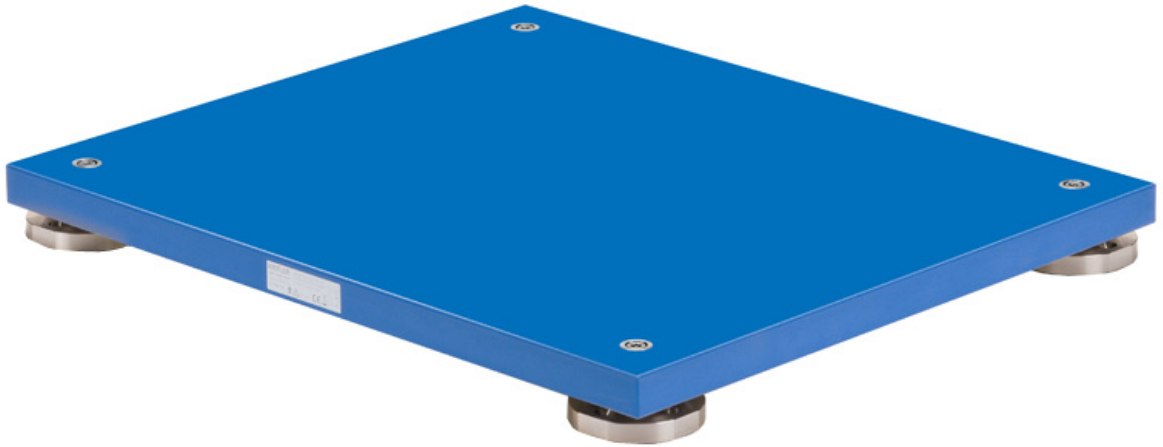
www.novel.de

KISTLER

measure. analyze. innovate.

DEXL LAB

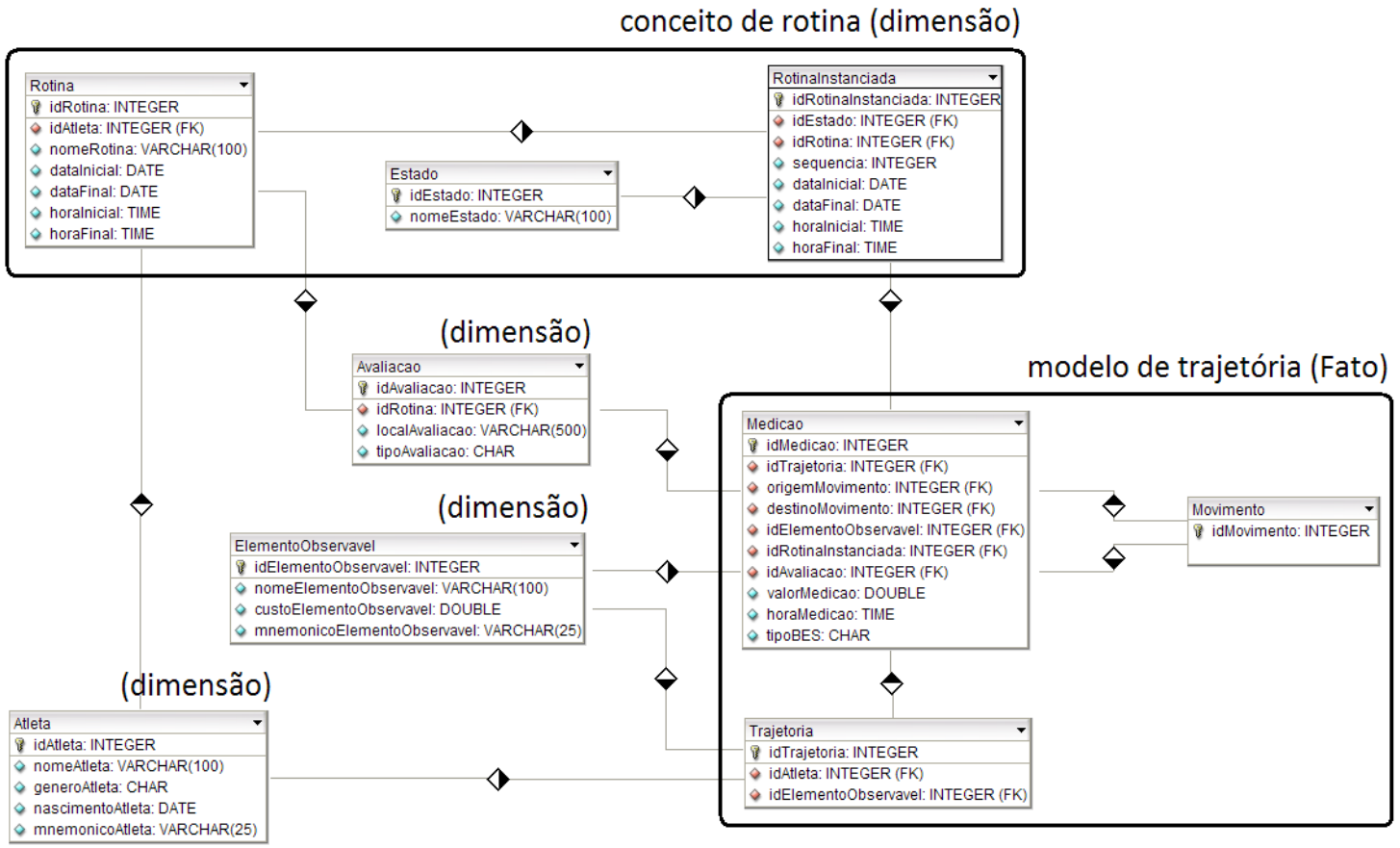
ME DATA LAB



Extração de Sangue

Espectrometria de Massa

Modelo Estrela para Análise de Dados



Relação: Alvo –Medições

Resultado da avaliação do atleta amador: Renato Pinheiro Oliveira

Alvo

Evidências

Hipóteses

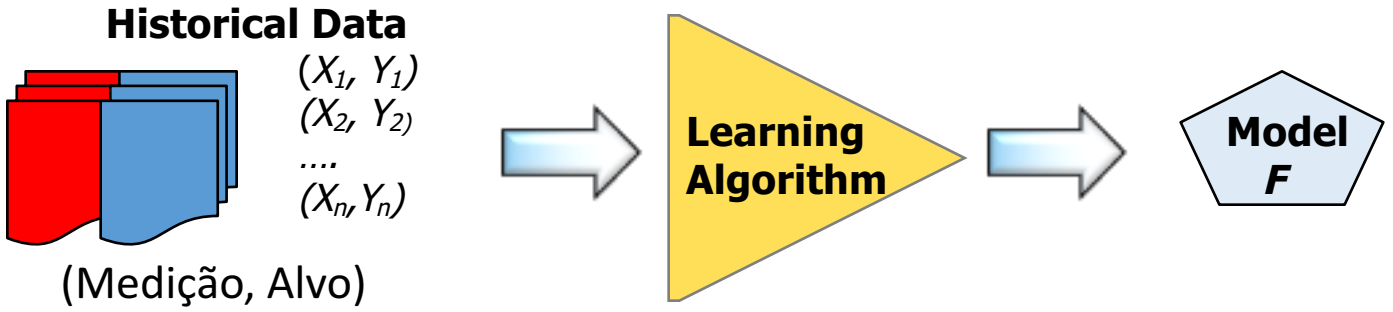
ESPORTE	OBJETIVO	PRAZO	#	DEPTO.	EVIDÊNCIAS	HIPÓTESES		SUGESTÕES
MARATONA	REDUÇÃO DO TEMPO EM 2 MINUTOS	1 ANO	1	BIOQUIMICA	Análise do nível de CK, GamaGT	Aumentar Treinamento		Implementação de práticas nutricionais suplementares
					Queda da concentração de bicarbonato	Aumentar reserva de bicarbonato		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				

Abordagem

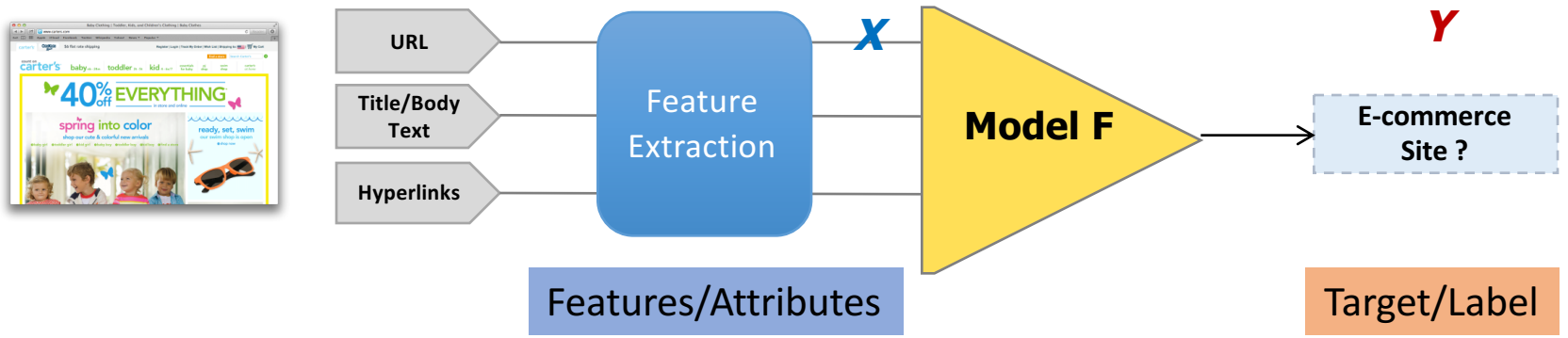
- Identificar atributos relevantes
  - Avaliação estatísticas de correlação entre alvos e medições
- Definir grupos de atributos correlacionados
  - Discretização de atributos (SAX)
  - Determinação de grau de Independência entre Informações Redundante [Wong et al 1976]
  - Uso de algoritmos de clusterização (k-means)
- Estabelecer relações probabilísticas de causalidade
  - Modelo Grafo Probabilístico -> Causal Bayesian Networks
- Identificar elementos a sofrerem interferência

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 $F(X)$

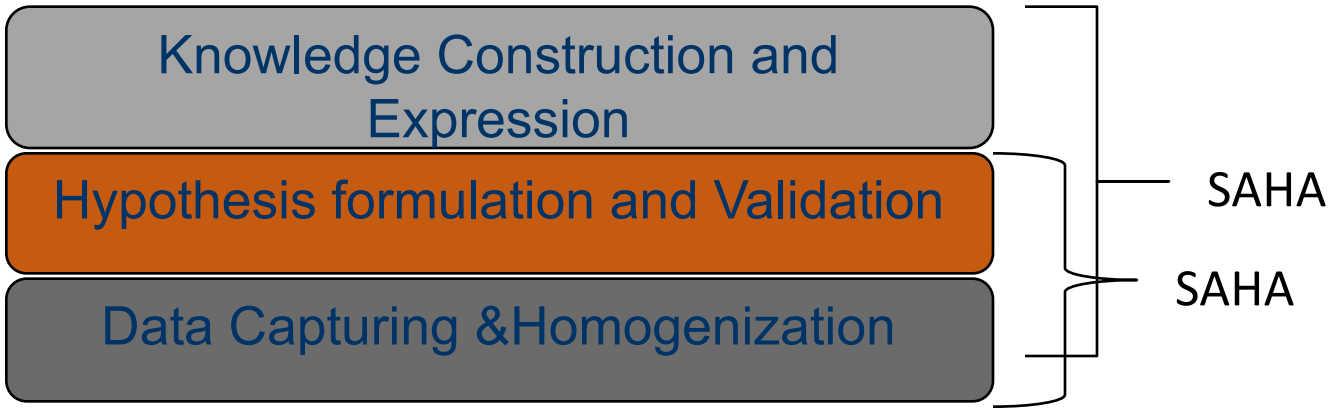


- **Inverted Problem:** Given a Y determine relevant X using $F^{-1}(Y)$

SAHA – Sistema de Apoio Holistico ao Atleta

SAHA: Apoio à análise integrada de dados de atletas de alto rendimento

DEXL LAB
EXTREME DATA LAB



XL LAB
ME DATA LAB

- Início
- Atleta/Equipe
- Elementos observáveis
- Métricas
- Rotina
- Coleta de dados
- **Gráficos**
  - **Análise de trajetória básica (atleta x exame)**
  - Média aritmética e desvio padrao (modalidade x exame x estado)
  - Análise de trajetórias ascendentes e descendentes
  - Análise de trajetórias com extremidades maiores ou menores
  - Análise de máximo e mínimos em uma trajetoria
  - Busca trajetórias pertencentes a intervalos máximos e mínimos

Gráficos

Atleta: Joao

Elemento observável: Colesterol

Data Inicial: 16/03/2015

Data Fim: 23/03/2015

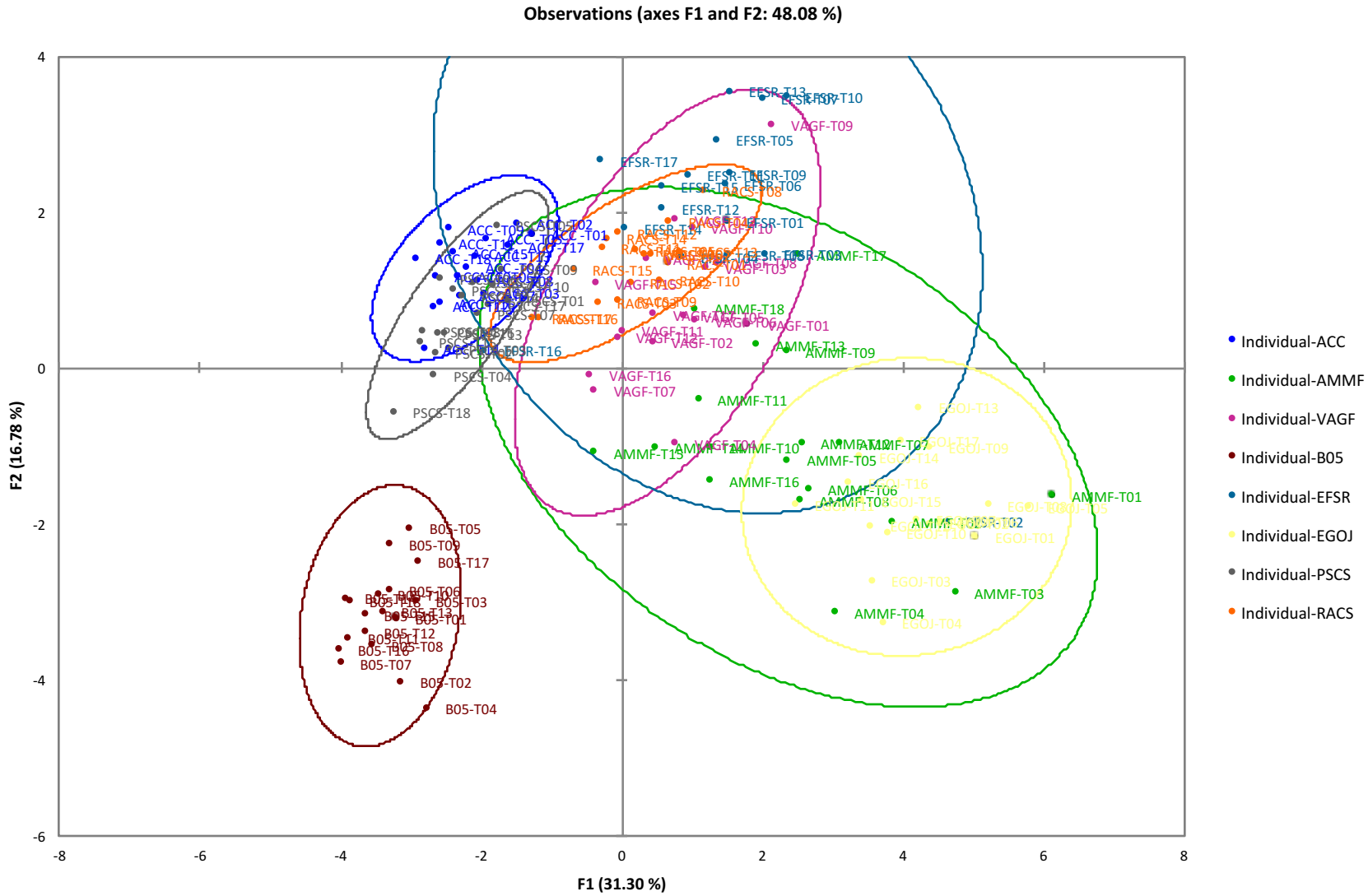
Estados:

- [x] jejum
- [ ] Corrida1
- [x] Pos Treino
- [ ] Repouso1
- [ ] 12321
- [x] Corrida
- [ ] gfdfgfdgdfg
- [ ] Corrida 100m
- [ ] 12321

Gerar gráfico

Gráfico 1 | Gráfico 2 | Gráfico 3 | Gráfico 4

PCA: F1 *vs* F2

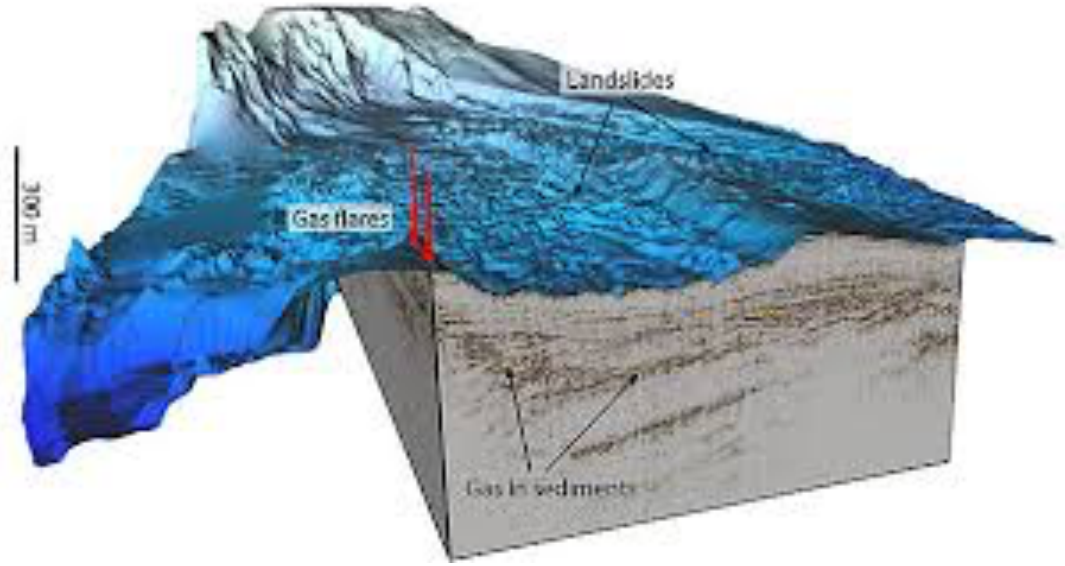


DEXL LAB
EXTREME DATA LAB

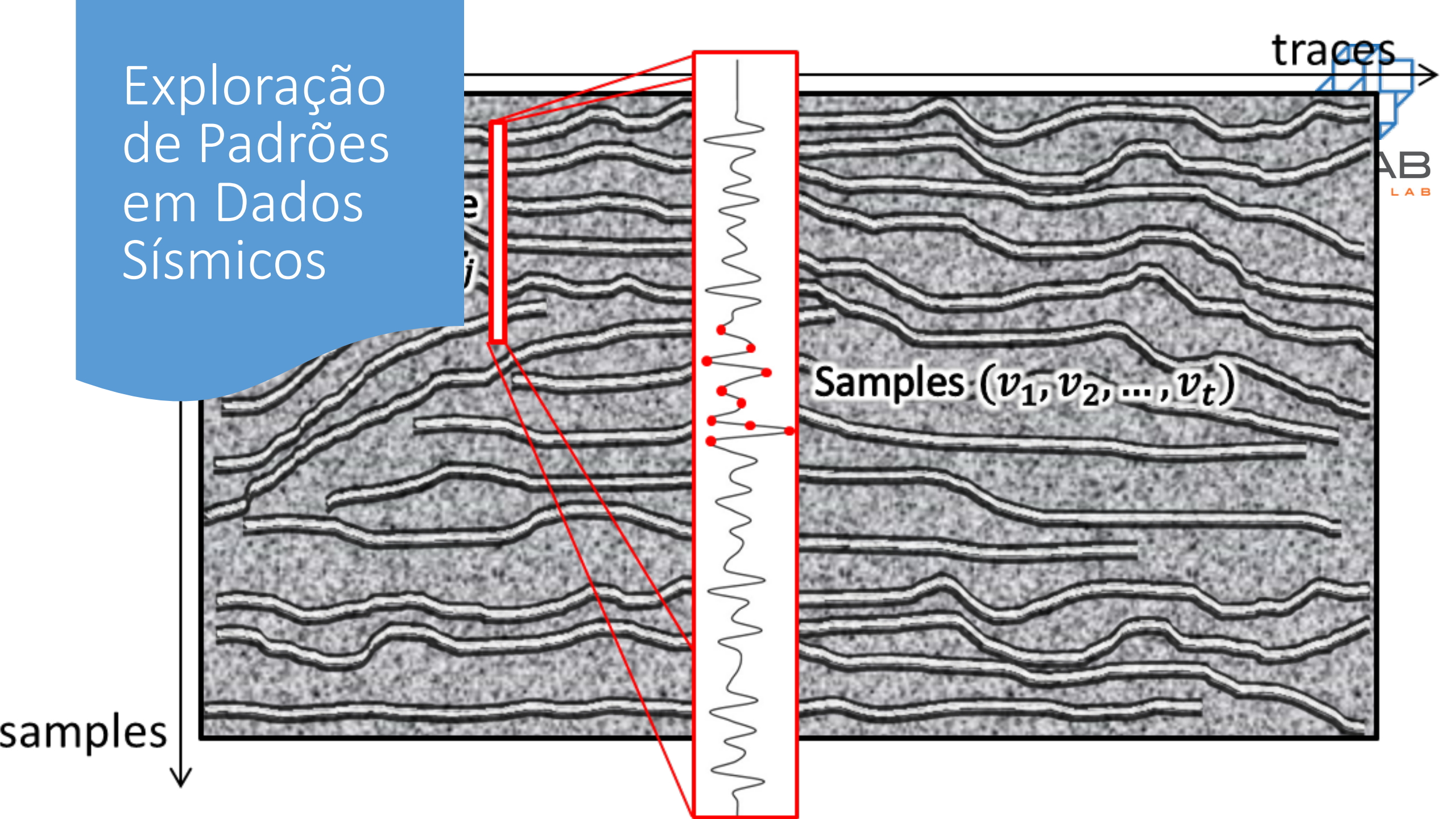
LBP
Laboratório de Bioquímica de Proteínas
UNIVERSIDADE FEDERAL DO ESTADO DO RIO DE JANEIRO

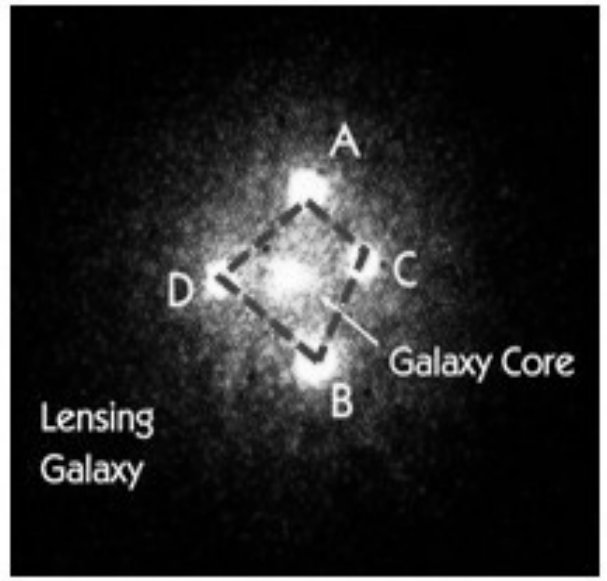
Séries espaço-temporais

DEXL LAB
EXTREME DATA LAB

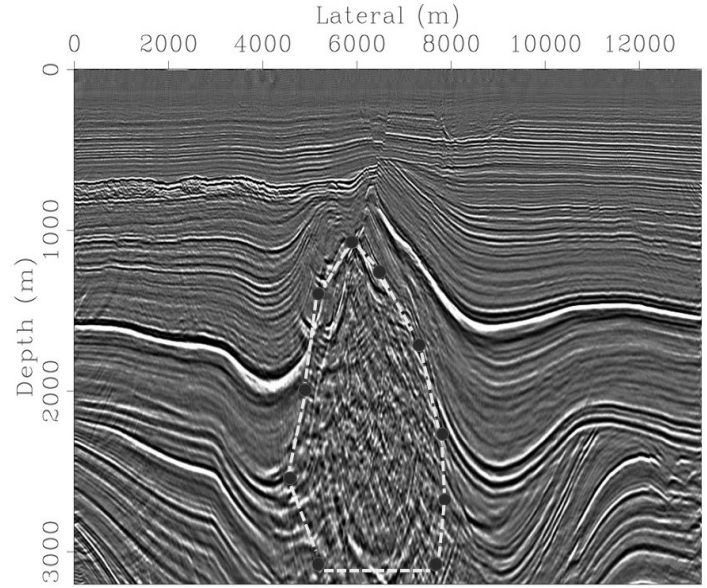


Exploração de Padrões em Dados Sísmicos

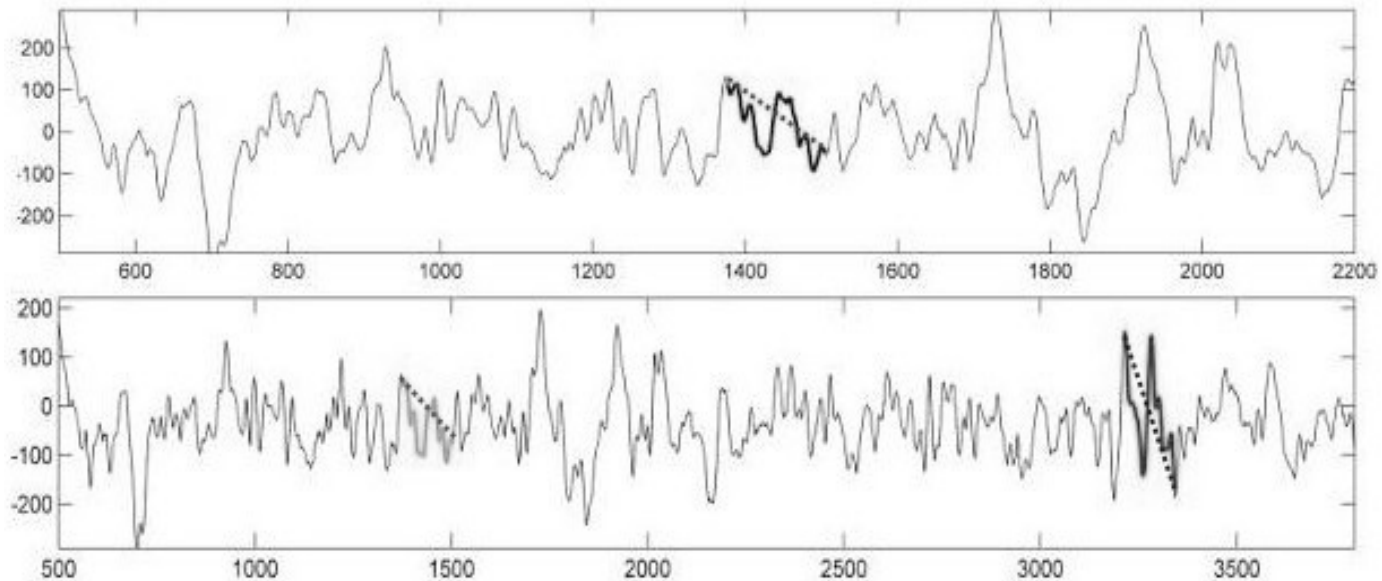




(a)



(b)



(c)

DEXL LAB
EXTREME DATA LAB

Identificação de padrões em Big Data

- Tratamento dos dados
  - Discretização/normalização
  - Indexação
  - Distribuição
- Análise
  - Algoritmos de análise de padrões em séries temporais
  - Configuração / análise paramétrica
  - Teste e avaliação
- Ranking de soluções
  - Métrica de comparação entre soluções

Comentários Finais

- Ciência de Dados
  - Uma nova área multi-disciplinar
  - Ciência da Computação como um dos alicerces na nova ciência
- O processo de investigação carece de sistemas de apoio
  - Ambiente com várias alternativas mas alguns pilares já estão mais sedimentados
- A importância de trabalhar em problemas reais com dados disponíveis
  - Não invente o seu problema!!!
- Os resultados pecisam ser validados e interpretados
  - Obtenha e estabeleça um *golden stardard*
  - Estabeleça o critério de avaliação para predições, principalmente sobre dados novos
- Modelos só são bons quando coincidem com a interpretação que se pretende extrair dos dados
- Jornada Em Ciência de Dados, Fev/2016

Um Excelente Início de Curso !!!

Obrigado !☺

Fabio Porto (fporto@lncc.br)

DEXL LAB
EXTREME DATA LAB