ciência de dados
+ computação inteligente
==
ciência da computação
+ dados inteligentes

Wagner Arbex
wagner.arbex@embrapa.br
wagner.arbex@ufjf.edu.br

Embrapa Gado de Leite

ICMC/USP — São Carlos — 25/8/2016

Computação científica...

... é também chamada *e-science*, *ciência computacional*, *ciência intensiva* etc. e não deve ser confundida com ciência da computação;

... é a aplicação de modelagem matemática e computacional em problemas científicos e tecnológicos;

... promove a computação massiva e/ou complexa, que são “novos” paradigmas da pesquisa científica;

Computação científica...

All models are wrong, but some are useful (George Box, 197[8])

All models are wrong, and increasingly you can succeed without them (Peter Norvig, 2008)

Ciência de dados...

WIRED MAGAZINE: 16.07

SCIENCE : DISCOVERIES

The End of Theory: The Data Deluge Makes the Scientific Method Obsolete

By Chris Anderson 06.23.08

"All models are wrong, but some are useful."

So proclaimed statistician George Box 30 years ago, and he was right. But what choice did we have? Only models, from cosmological equations to theories of human behavior, seemed to be able to consistently, if imperfectly, explain the world around us. Until now. Today companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all.

Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age.

Computação científica...

COMPUTAÇÃO CIENTÍFICA...

...is like youthful 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 ...

Ciência de dados...

MIT Technology Review

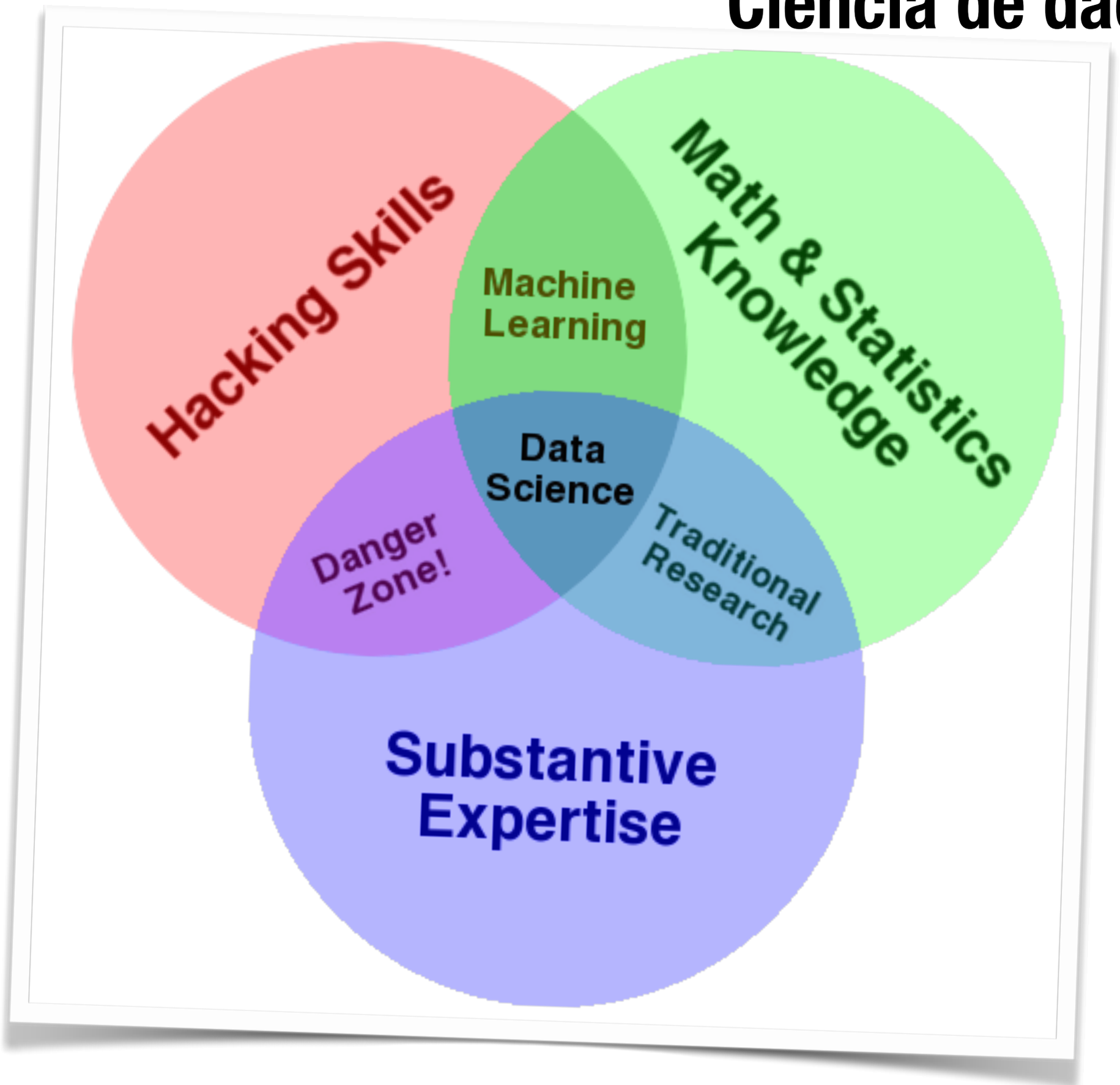
In a Data Deluge, Companies Seek to Fill a New Role

By Jessica Leber on May 22, 2013

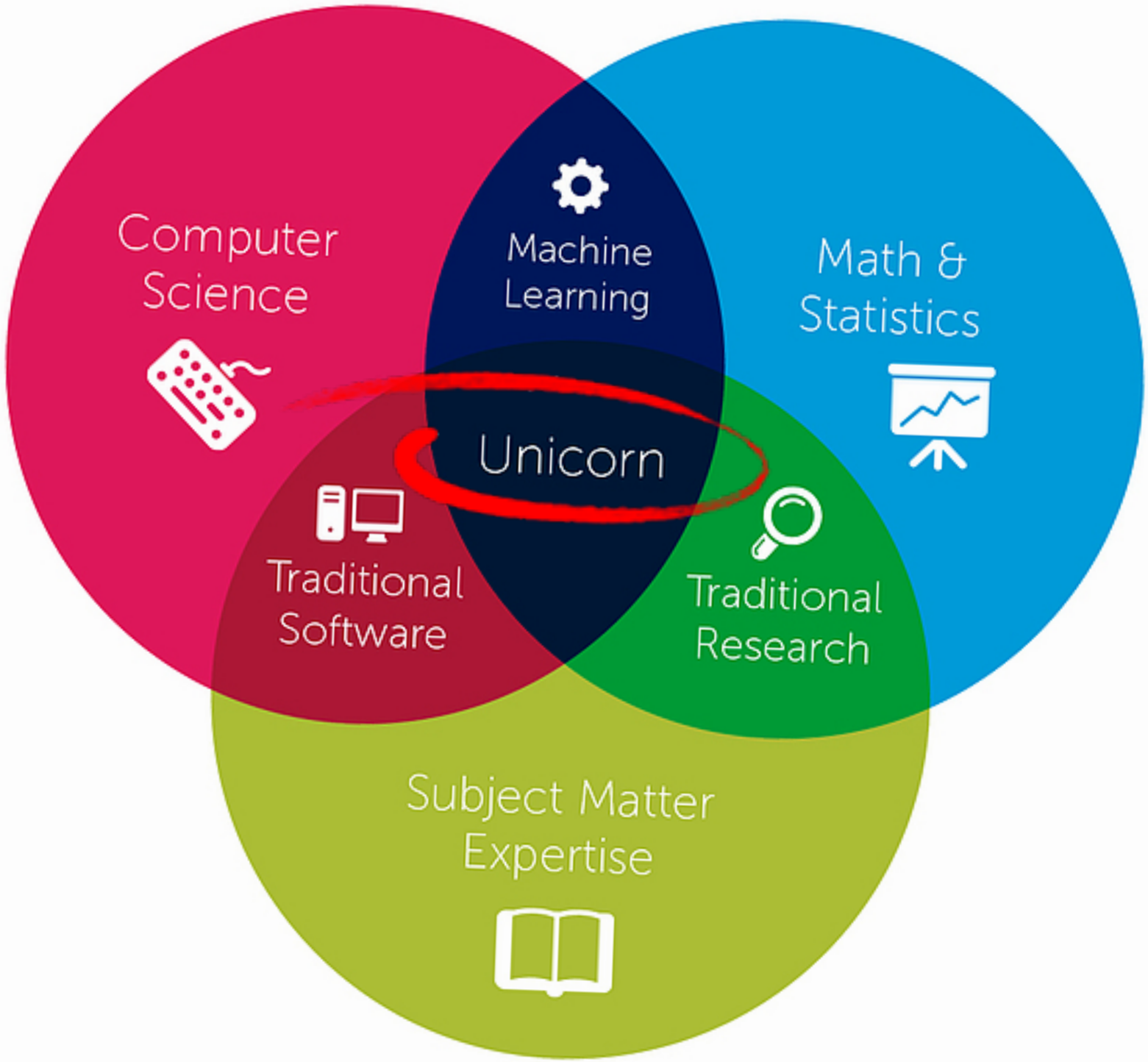
The job description "data scientist" didn't exist five years ago. No one advertised for an expert in data science, and you couldn't go to school to specialize in the field. Today, companies are fighting to recruit these specialists, courses on how to become one are popping up at many universities, and the *Harvard Business Review* even proclaimed that data scientist is the "sexiest" job of the 21st century.

Data scientists take huge amounts of data and attempt to pull useful information out. The job combines statistics and programming to identify sometimes subtle factors that can have a big impact on a company's bottom line, from whether a person will click on a certain type of ad to whether a new chemical will be toxic in the human body.

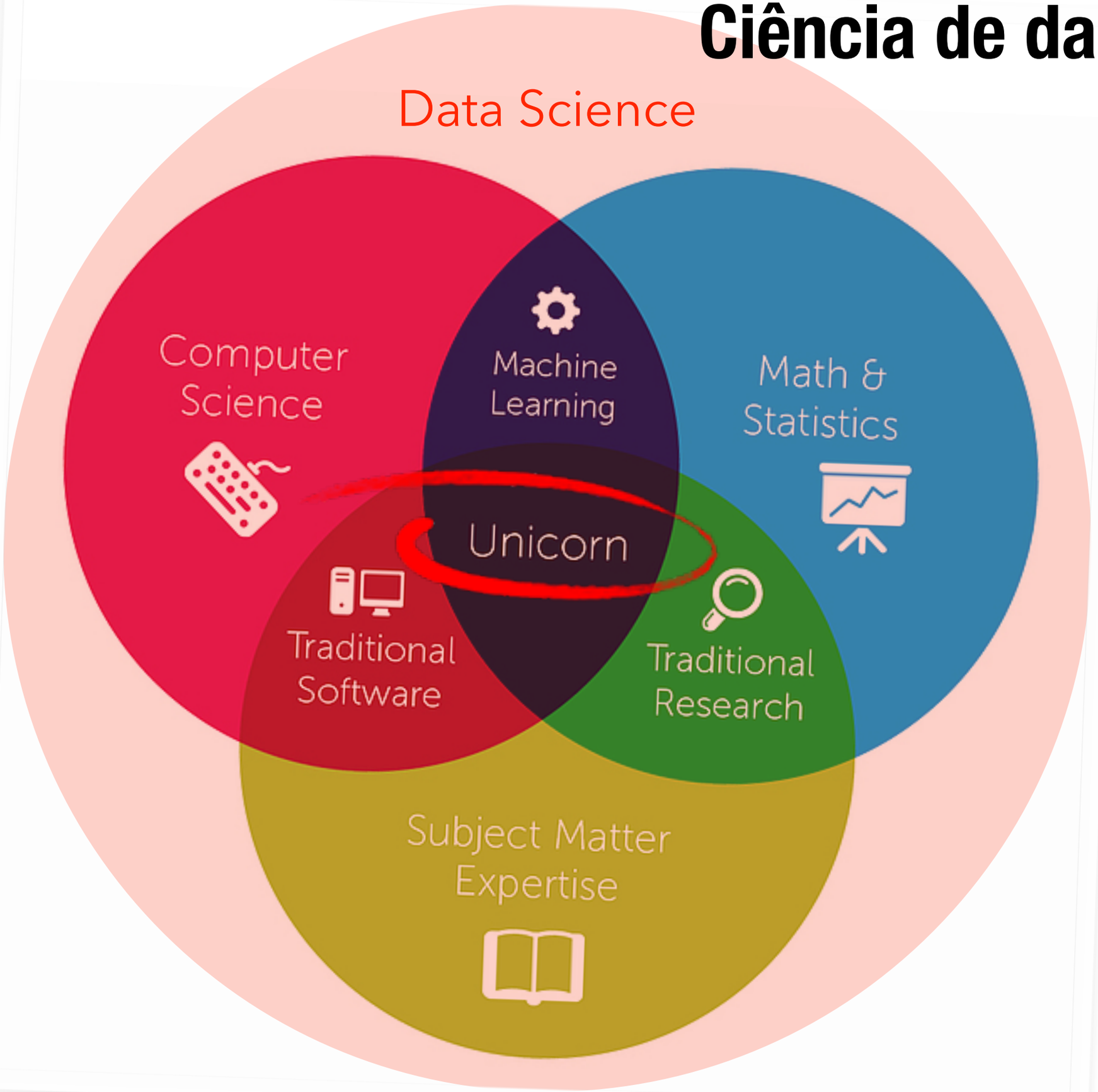
Ciência de dados…



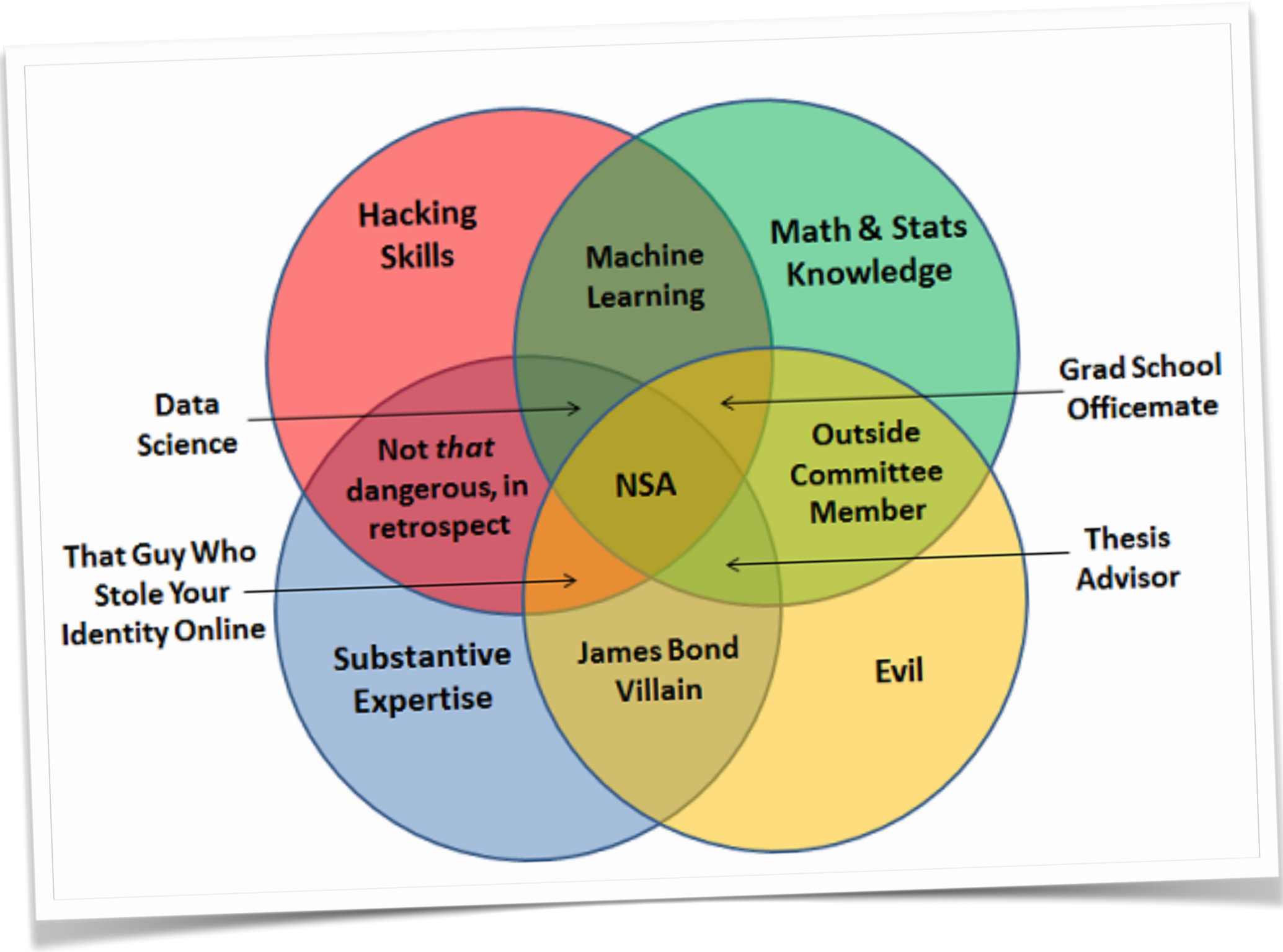
Ciência de dados…



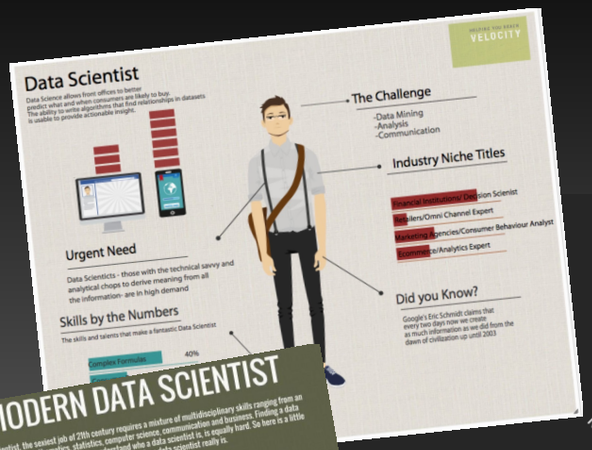
Ciência de dados…

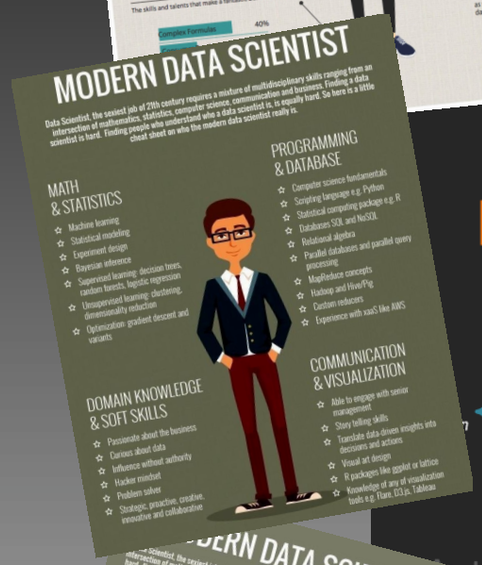


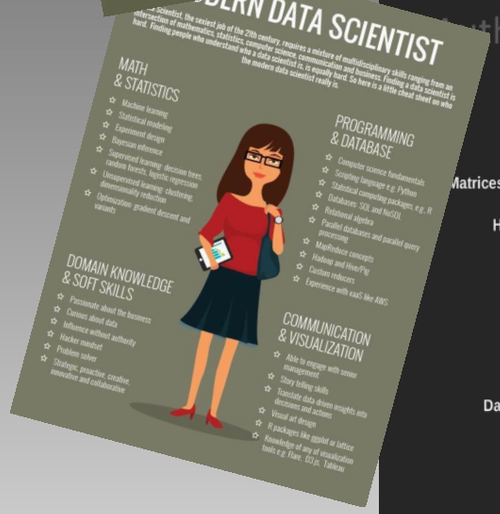
Ciência de dados…



Cientista de dados, The Unicorn







1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%

2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
15%
Factor Analysis
Install Pkgs

3. Programming
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Lists
Factors
Arrays
Matrices
Vectors
Variables
Expressions
R Basics
R Setup R Studio
Python Basics
Working in Excel
Rapid Miner
IBM SPSS

4. Machine Learning
Named Entity Recognition
Text Analysis
Corpus
Clustering
Hierarchical Clustering
K-means Clustering
Term Document Matrix
UIMA
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Perceptron
Linear Regression
Ranking
Logistic Regression
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
Vocabulary Mapping
Classify Text
50%
Feature Extraction
Using Mahout
Using Weka
Using NLTK
Classification
K-Nearest Neighbor
Naïve Bayes Classifiers
Boosting
Decision Trees
Classification Rate
Trees & Classification
Bias & Variance
Overfitting
Lift
Prediction
Classifier
Training & Test Data
Concepts, Inputs & Attributes
Unsupervised Learning
Supervised Learning
Categorical Var
Numerical Var
What is ML?
30%
Euclidean Distance
Least² Fit
Causation
Pearson Coeff
Correlation

5. Text Mining / NLP

6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
40%
Tableau
IBM ManyEyes
InfoVis
D3.js

7. Big Data
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe : For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Name & Data Nodes
Setup Hadoop (IBM / Cloudera / HortonWorks)
Data Replication Principles
HDFS
Hadoop Components
Map Reduce Fundamentals
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%

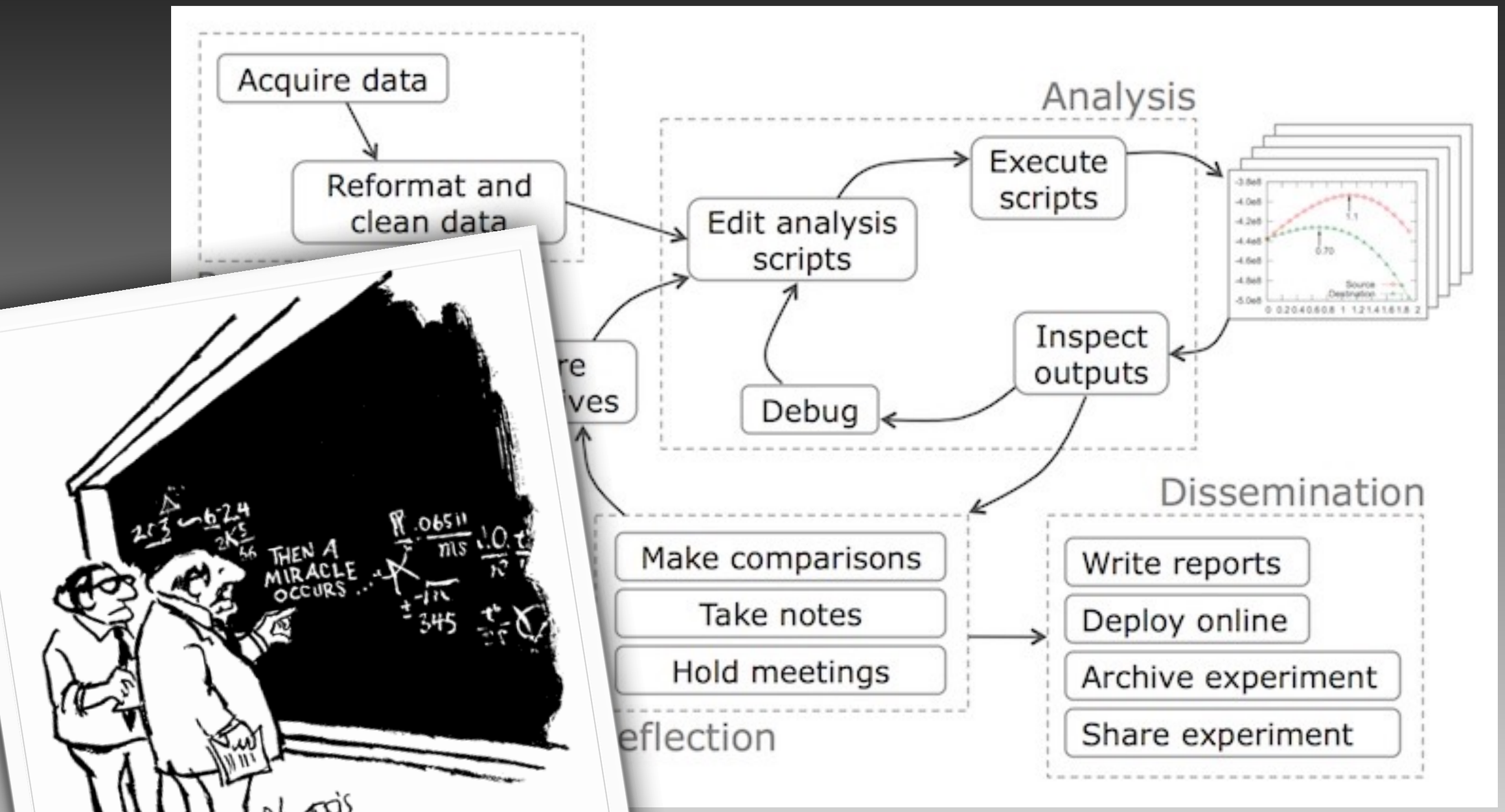
8. Data Ingestion
Sampling
Using ETL
How much Data?
Google OpenRefine
Data Survey
80%
Stratified Sampling
Principal Component Analysis
Transformation & Enrichment
Data Fusion
Data Integration
Data Sources & Acquisition
Data Discovery
Summary of Data Formats

9. Data Munging
Denoising
Feature Extraction
Binning Sparse Values
Unbiased Estimators
Handling Missing Values
Data Scrubbing
Normalization
Dimensionality & Numerosity Reduction

10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%

Ciência de dados…

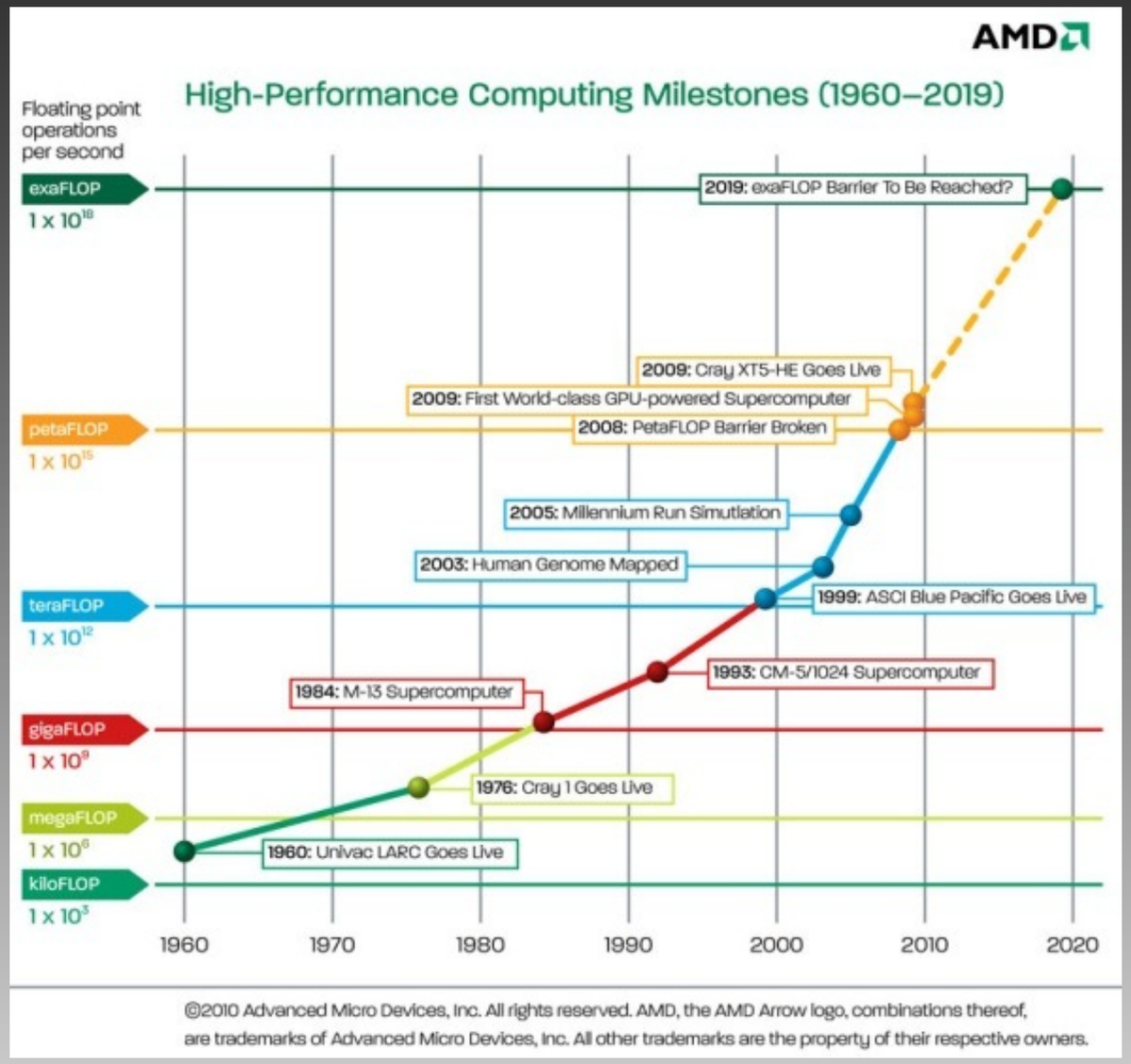
Ciência de dados…



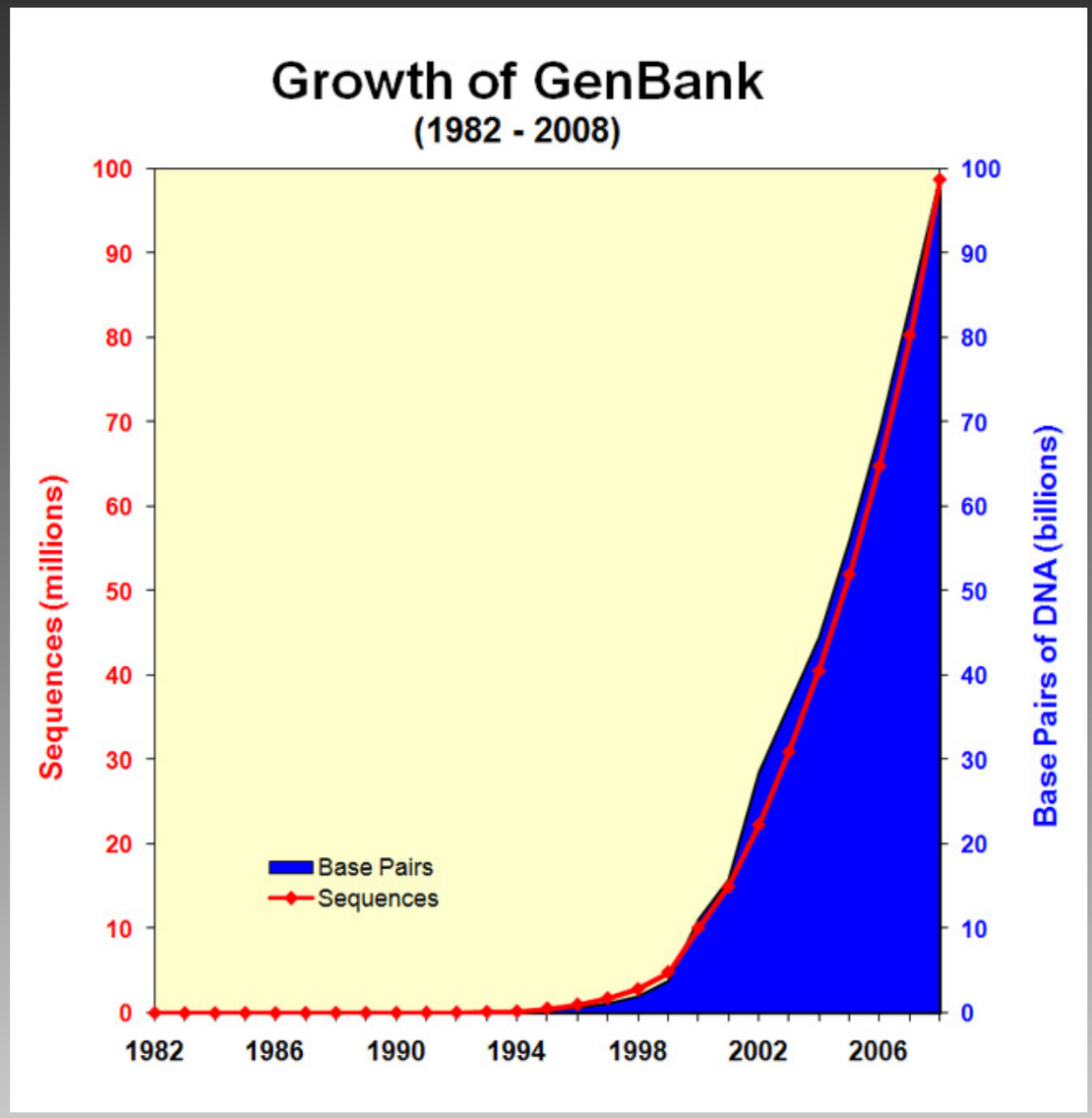
Science Workflow: Overview and Challenges

Mas, o que eu faço com isso?

Mas, o que eu faço com isso?

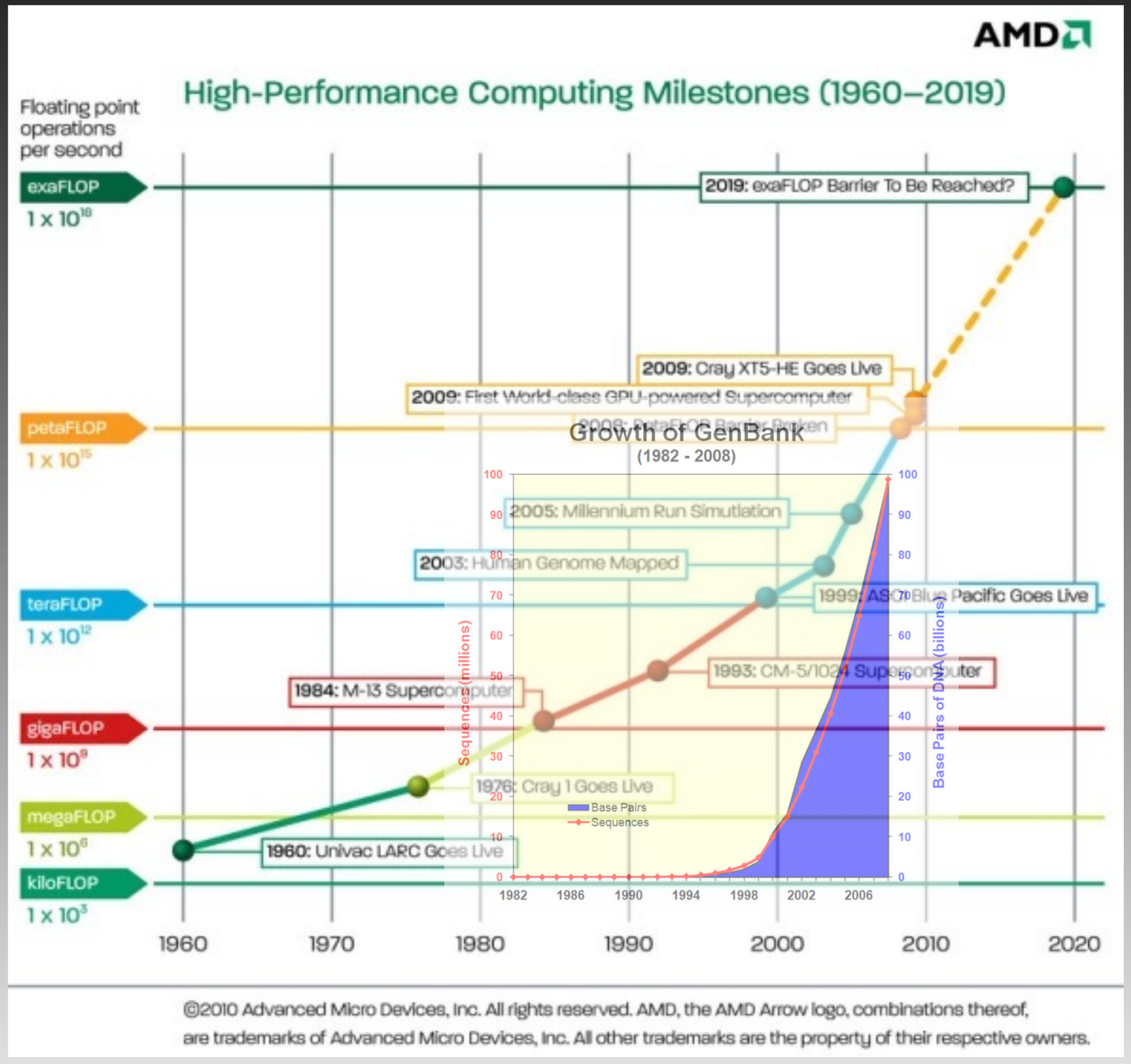


Fonte: AMD (2010)

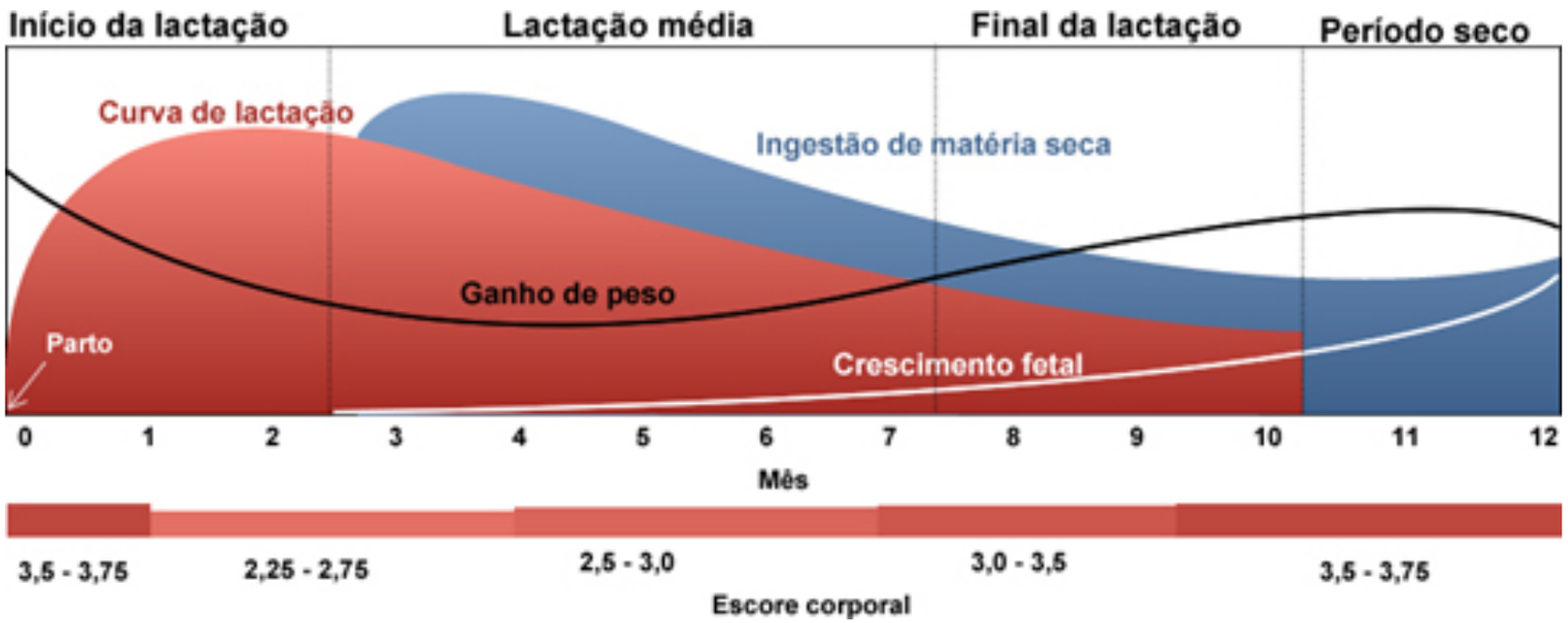


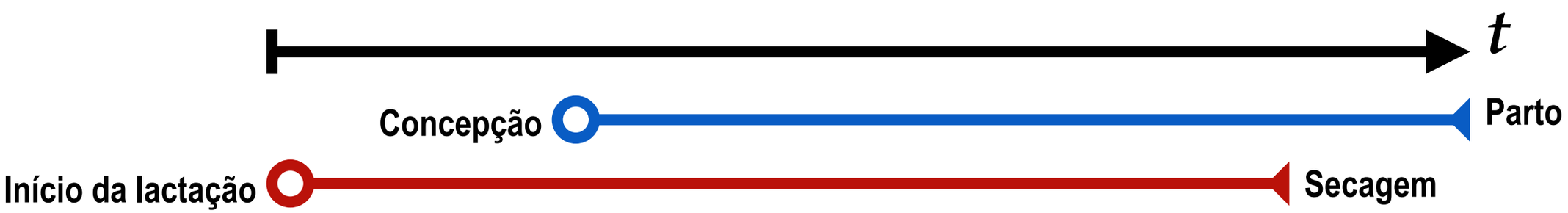
Fonte: NCBI (2008)

Mas, o que eu faço com isso?

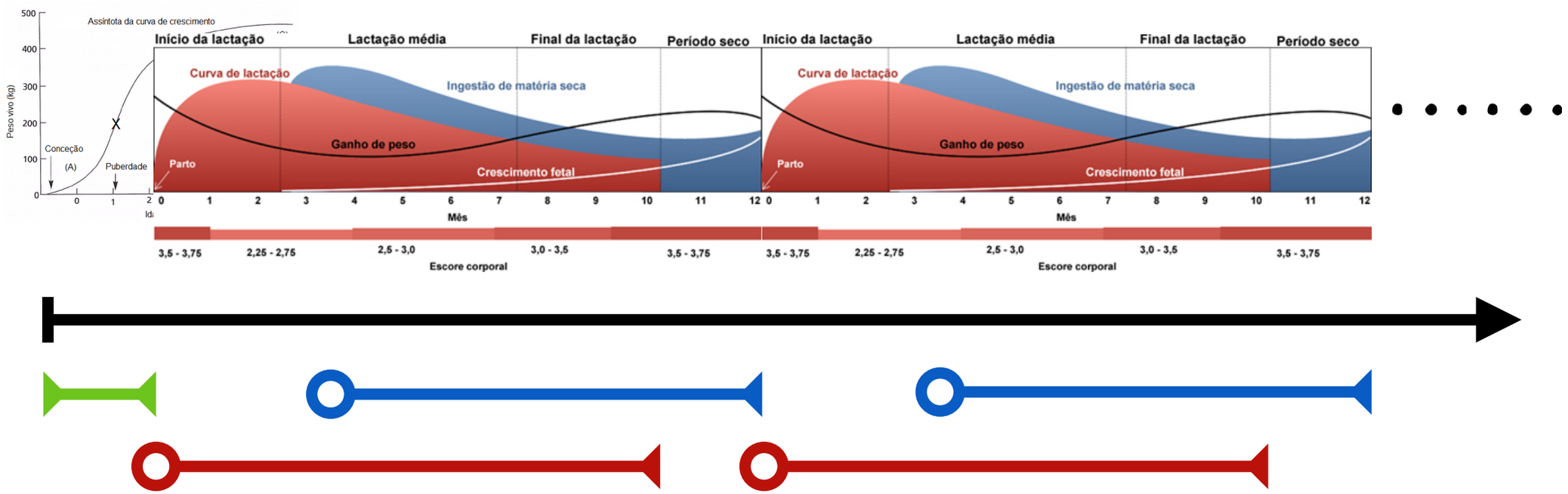


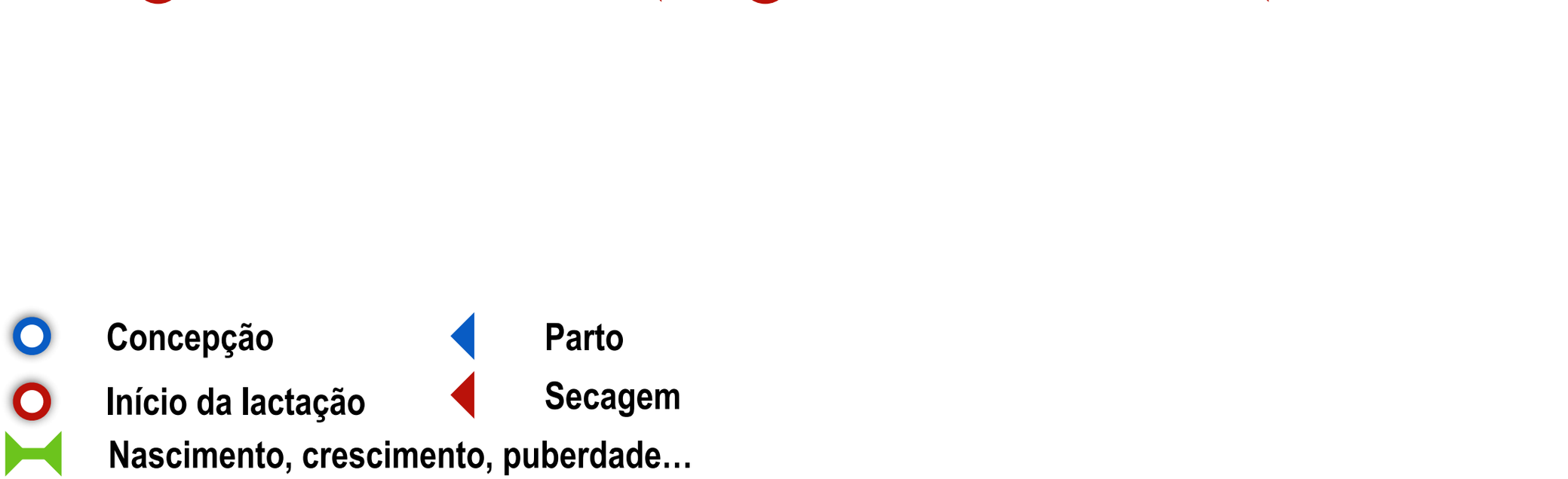
Aplicações


Início da lactação
Lactação média
Final da lactação
Período seco
Curva de lactação
Ingestão de matéria seca
Ganho de peso
Parto
Crescimento fetal
0 1 2 3 4 5 6 7 8 9 10 11 12
Mês
3,5 - 3,75
2,25 - 2,75
2,5 - 3,0
3,0 - 3,5
3,5 - 3,75
Escore corporal

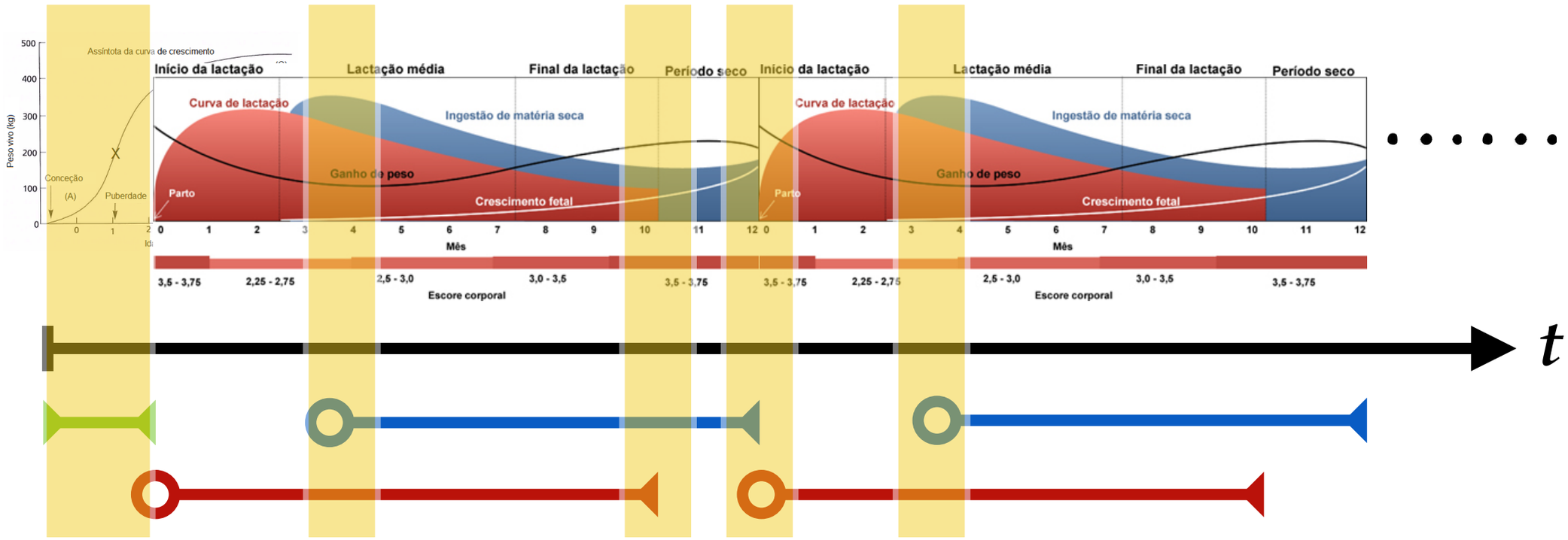

t
Concepção
Parto
Início da lactação
Secagem

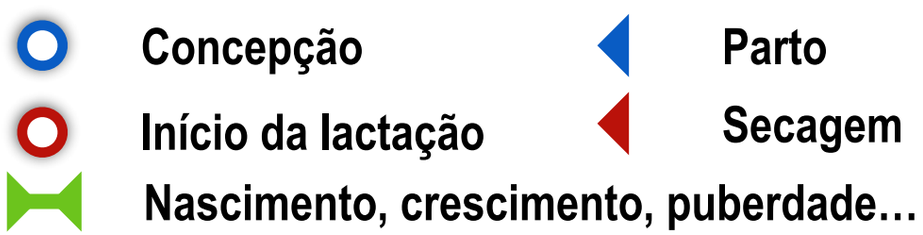
Aplicações


Assíntota da curva de crescimento
Peso vivo (kg)
500
400
300
200
100
0
Concepção
(A)
Puberdade
Início da lactação
Lactação média
Final da lactação
Período seco
Curva de lactação
Ingestão de matéria seca
Ganho de peso
Crescimento fetal
Parto
0 1 2 3 4 5 6 7 8 9 10 11 12
Mês
3,5 - 3,75
2,25 - 2,75
2,5 - 3,0
3,0 - 3,5
3,5 - 3,75
Escore corporal
. . . . . .


Concepção
Início da lactação
Nascimento, crescimento, puberdade...
Parto
Secagem

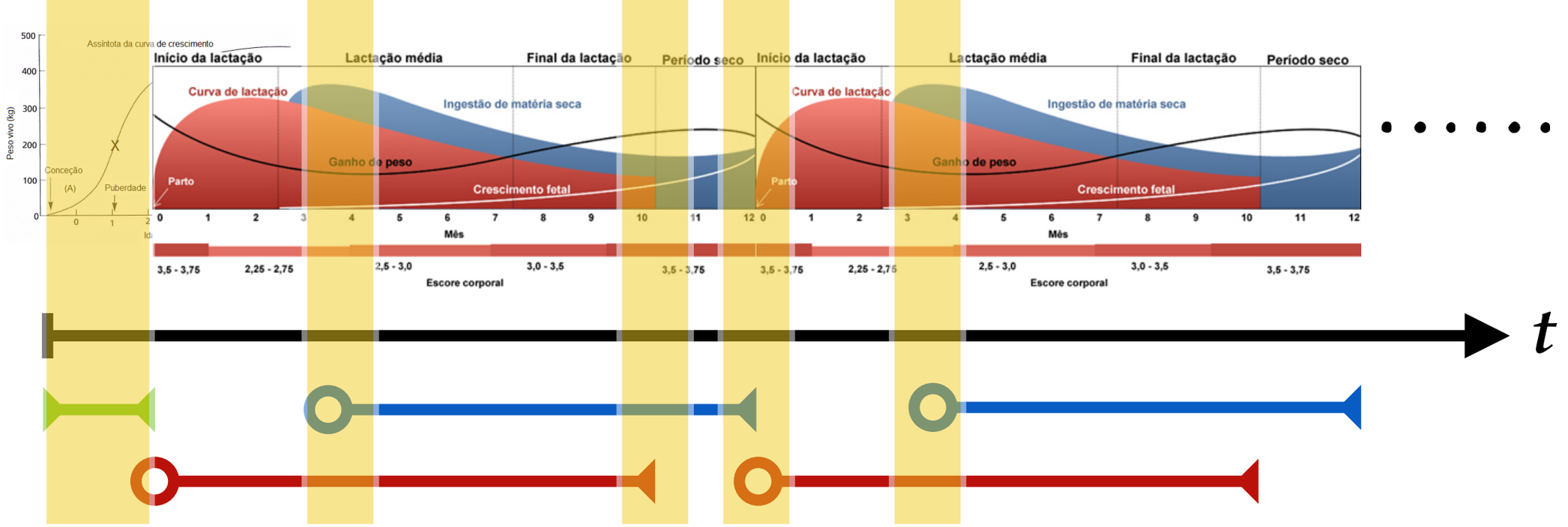
Aplicações


Assíntota da curva de crescimento
Peso vivo (kg)
500
400
300
200
100
0
Concepção
(A)
Puberdade
Início da lactação
Lactação média
Final da lactação
Período seco
Curva de lactação
Ingestão de matéria seca
Ganho de peso
Parto
Crescimento fetal
0 1 2 3 4 5 6 7 8 9 10 11 12
Mês
3,5 - 3,75
2,25 - 2,75
2,5 - 3,0
3,0 - 3,5
Escore corporal
t


Concepção
Início da lactação
Nascimento, crescimento, puberdade...
Parto
Secagem

Aplicações

IoToMilk



Concepção

Início da lactação

Nascimento, crescimento, puberdade…

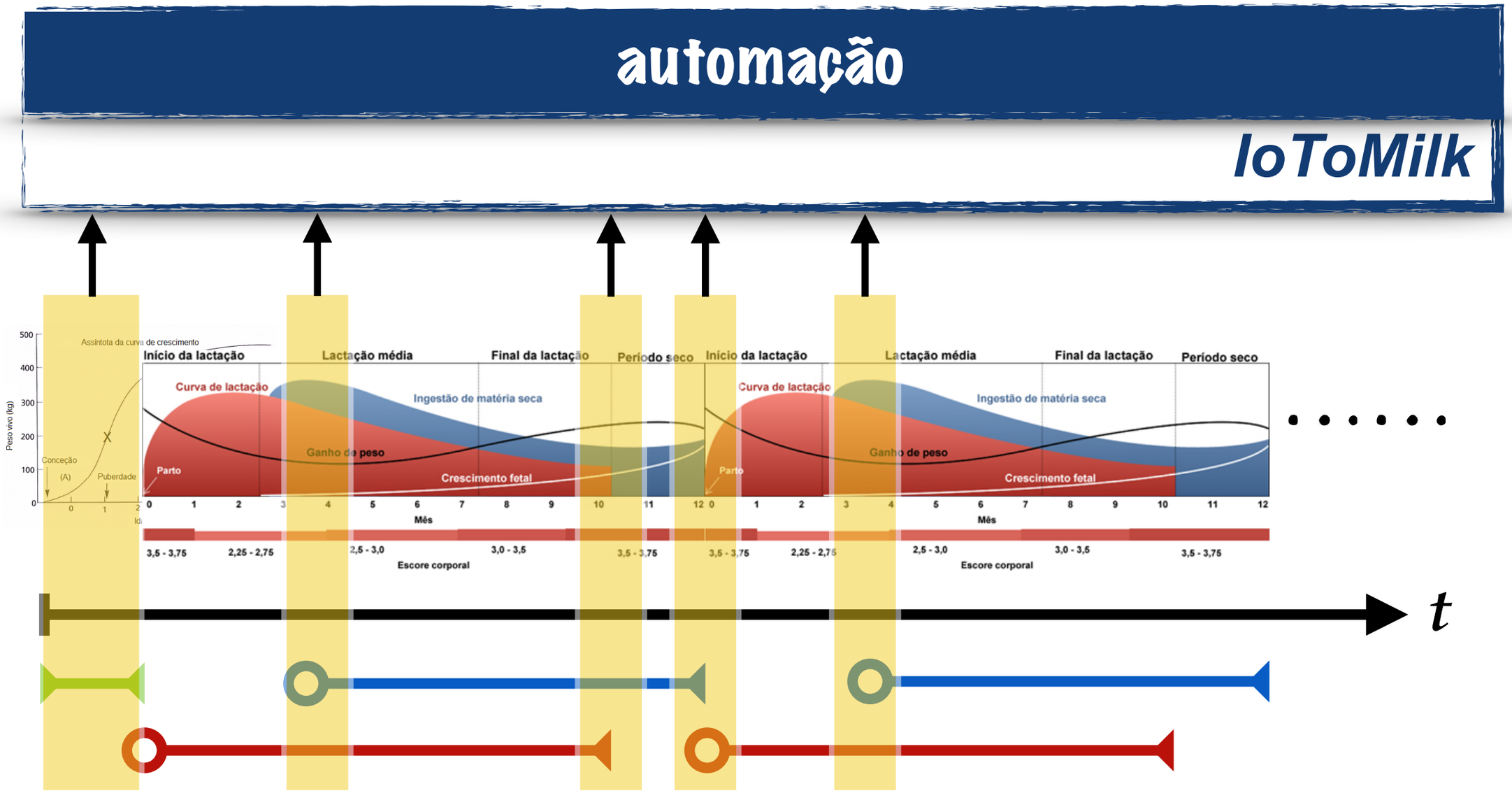
Parto

Secagem

Aplicações

automação

IoToMilk



Concepção

Início da lactação

Nascimento, crescimento, puberdade…

Parto

Secagem

Infraestrutura para ciência de dados

A Low-cost Infrastructure for Massive Storage of Phenotype Data for Dairy Cattle Genetic Improvement Programs

Wagner Arbex*,†, Caio dos Santos Borsato de Carvalho*,‡, Katia Santos*,
Vinicius Campista Brum§ and Marcos Vinícius Barbosa da Silva*
*Brazilian Agricultural Research Corporation (Embrapa)
Juiz de Fora, MG, Brazil
†Correspondent author – wagner.arbex@embrapa.br
‡National Council for Scientific and Technological Development (CNPq) grant
§Federal University of Juiz de Fora (UFJF)

***Abstract*—The activities and progress of genetic breeding were always related to computing – or rather, the availability of appropriate computational resources to the behavior of genetic improvement, because, either by cross breeding or selection, the data sets involved in genetic improvement research and development activities care both in quantity and in quality. The demand for computing resources in these programs is extensive and intense, because genetic genomic evaluation**

Add up to these aspects the need to interpret and understand the phenotype databases from a logical and effective structure for data storage and information retrieval.

Initiatives for data storage in genetic breeding programs are not unknown. Among others, the National Dairy Cattle Research Center (Embrapa Dairy Cattle) of the Brazilian Agricultural Research Corporation (Embrapa)

Banco de dados não convencionais

Genotypes_50K — -bash — 115×33

```
pc10122:Genotypes_50K arbex$ head -30 genotipo-HOL-56Kx577.txt



[Header]
BSGT Version     3.2.23
Processing Date 8/14/2008 11:44 AM
Content          USDA_Bovine_58K_271441_A.bpm
Num SNPs         56947
Total SNPs       56947
Num Samples      577
Total Samples    1630
[Data]
SNP Name         Sample ID       Allele1 - AB    Allele2 - AB
ARS-BFGL-BAC-10172       53998679        B       B
ARS-BFGL-BAC-1020        53998679        A       B
ARS-BFGL-BAC-10245       53998679        B       B
ARS-BFGL-BAC-10345       53998679        A       B
ARS-BFGL-BAC-10365       53998679        -       -
ARS-BFGL-BAC-10375       53998679        B       B
ARS-BFGL-BAC-10591       53998679        A       B
ARS-BFGL-BAC-10793       53998679        B       B
ARS-BFGL-BAC-10867       53998679        A       B
ARS-BFGL-BAC-10919       53998679        B       B
ARS-BFGL-BAC-10951       53998679        -       -
ARS-BFGL-BAC-10952       53998679        A       A
ARS-BFGL-BAC-10960       53998679        B       B
AR                       53998679        A       A
AF                                       -
AF
AI
p
```

X-meeting • November 2015 • USP Genomics

Storage and recovery of dairy cattle genotype data from the data science approach

Rennan Silva, Fernanda Almeida, Wagner Arbex

Juiz de Fora, Embrapa Dairy Cattle

Abstract

s of the identification of molecular markers, which may va
ually, the records that identify these markers are stored in
ormations about each individual, like the animal number a
larker. There are different patterns to present the genotyp
orm chosen for the job. Usually, the data gathered from
s. Analytical tools are not used to treat this kind of data be
e. There are several limitations on relational databases when
goal of this abstract is to propose a way to store the results
NP) and allow this information to be accessed from a middl

Avaliação do desempenho relativo de SGBDs NoSQL para arquivos de genótipos

Vinícius Junqueira Schettino[1], **Arthur Lorenzi Almeida**[1], **Leojayme Rodrigues Manso Silva**[1], **Wagner Arbex**[1,2,*]

[1]Universidade Federal de Juiz de Fora – UFJF
Dep. de Ciência da Computação – DCC
Campus Universitário, 36.036-900, Juiz de Fora, MG, Brasil

[2]Empresa Brasileira da Pesquisa Agropecuária – Embrapa
R. Eugênio do Nascimento, 610, 36.038-330, Juiz de Fora, MG, Brasil

schettino.vinicius@...

Seleção de variáveis e tomada de decisão…

Seleção de variáveis e tomada de decisão…

4

Metodologia para seleção de marcadores com máquina de vetores de suporte com regressão

Fabrízzio Condé de Oliveira
Fernanda Nascimento Almeida
Fabyano Fonseca e Silva
Marcos Vinícius Gualberto Barbosa da Silva
Carlos Cristiano Hasenclever Borges
Wagner Arbex

Seleção de variáveis e tomada de decisão…

NeuroSNP: Tool to Filter SNPs in Whole Genomic DNA

B. Zonovelli[1], **C. C. H. Borges**[1], **and W. A. Arbex**[2]
[1]Federal University of Juiz de Fora, Juiz de Fora, MG, Brazil
[2]Brazilian Agricultural Research Corporation, Juiz de Fora, MG, Brazil

Abstract— *The classification task the punctual differences in data f found when comparing the genomes f uals is complex. In dr higher than a final ver noise, this can be in th this noise difficult the p ences how being single not. The SNPs are speci of aligned sequences a of genetic variation. To the technique of the neu whole genome the differ*

Keywords: Bioinformatics
chine Learning, Computatio

Decision Support in Attribute Selection with Machine Learning Approach

Wagner Arbex
Brazilian Agricultural Research Corporation — Embrapa
Juiz de Fora, MG, Brazil
wagner.arbex@embrapa.br

Fabrízzio Condé de Oliveira
Federal University of Juiz de Fora — UFJF
Juiz de Fora, MG, Brazil

Fabyano Fonseca e Silva
Federal University of Viçosa — UFV
Viçosa, MG, Brazil

Luis Varona
University of Zaragoza — UNIZAR
Zaragoza, Spain

Rui da Silva Verneque

Seleção de variáveis e tomada de decisão…

de Oliveira *et al. BMC Genomics* 2014, **15**(Suppl 7):S4
http://www.biomedcentral.com/1471-2164/15/S7/S4

BMC Genomics

RESEARCH

Open Access

SNPs selection using support vector regression and genetic algorithms in GWAS

Fabrízzio Condé de Oliveira[1], Carlos Cristiano Hasenclever Borges[1], Fernanda Nascimento Almeida[2,4], Fabyano Fonseca e Silva[3], Rui da Silva Verneque[4], Marcos Vinicius GB da Silva[4], Wagner Arbex[1,4*]

From 9th International Conference of the Brazilian Association for Bioinformatics and Computational Biology (X-Meeting 2013)
Recife, Brazil. 3-6 November 2013

Abstract

Introduction: This paper proposes a new methodology to simultaneously select the most relevant SNPs markers for the characterization of any measurable phenotype described by a continuous variable using Support Vector Regression with Pearson Universal kernel as fitness function of a binary genetic algorithm. The proposed methodology is multi-attribute towards considering several markers simultaneously to explain the phenotype and is based jointly on statistical tools, machine learning and computational intelligence.

Results: The suggested method has shown potential in the simulated database 1, with additive effects only, and real database. In this simulated database, with a total of 1,000 markers, and 7 with major effect on the phenotype and the other 993 SNPs representing the noise, the method identified 21 markers. Of this total, 5 are relevant SNPs between the 7 but 16 are false positives. In real database, initially with 50,752 SNPs, we have reduced to 3,073

Identificação e avaliação de padrões

Computational methods applied to identification of the Dairy Gir breed families

Gisele Silva, Tales Silva, Míria Bobó, Fernanda Almeida, Victor Mo
Soares, João Cláudio Panetto, Wa

Federal University of Juiz de Fora (UFJF
(E

Ab

The Embrapa Dairy Cattle and the Brazili
sociação Brasileira dos Criadores de Gir Leitei
Improvement National Program of the Dairy Gir
Leiteiro/PNMGL). Released in May 2015, the mos
published the results about 300 males, using mo
or contemporary thereof and more over 32,000 re
and contemporary. Among several sets of data, the
two databases (i) for store the certificate (registry

X-meeting • November 2015 • USP Genomics

Applications of graph theory for verify the family relationship for genetic evaluations through the Animal Model

Pedro Bittencourt, Fernanda Almeida, Wagner Arbex

Federal University of Juiz de Fora (UFJF), Brazilian Agricultural Research Corporation (Embrapa)

Abstract

Animal genetic improvement is an usual method for increasing the productivity of the herd. Therefore, selecting animals that are potentially ̈betterthan their peers and/or contemporary is important because the descendants of these animals will have enhanced characteristics. Genetic evaluations are used in animal genetic improvement programs to predict the potential genetic value of the animals and their PTAs (predicted transmitting abilities); however, to obtain good results, one needs as much information as possible about the individuals, their relatives, and their

ciência de dados
+ computação inteligente
==
ciência da computação
+ dados inteligentes

Wagner Arbex
wagner.arbex@embrapa.br
wagner.arbex@ufjf.edu.br

Embrapa Gado de Leite

ICMC/USP — São Carlos — 25/8/2016