Redes de Informação/Ciências da Informação

CONEXÃO DE INFORMAÇÕES

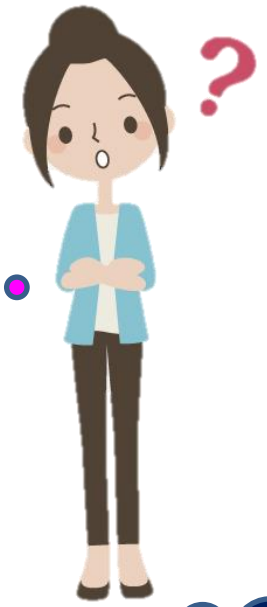
"MACHINE LEARNING?"

POR ONDE COMEÇAR?

Pós-doc – Iara Del Pilar Solar Diaz/ Genética e Melhoramento Animal/UFBA

Democratização do conhecimento?

Tipos de dados coletados



Dados Compartilhados x privatizados

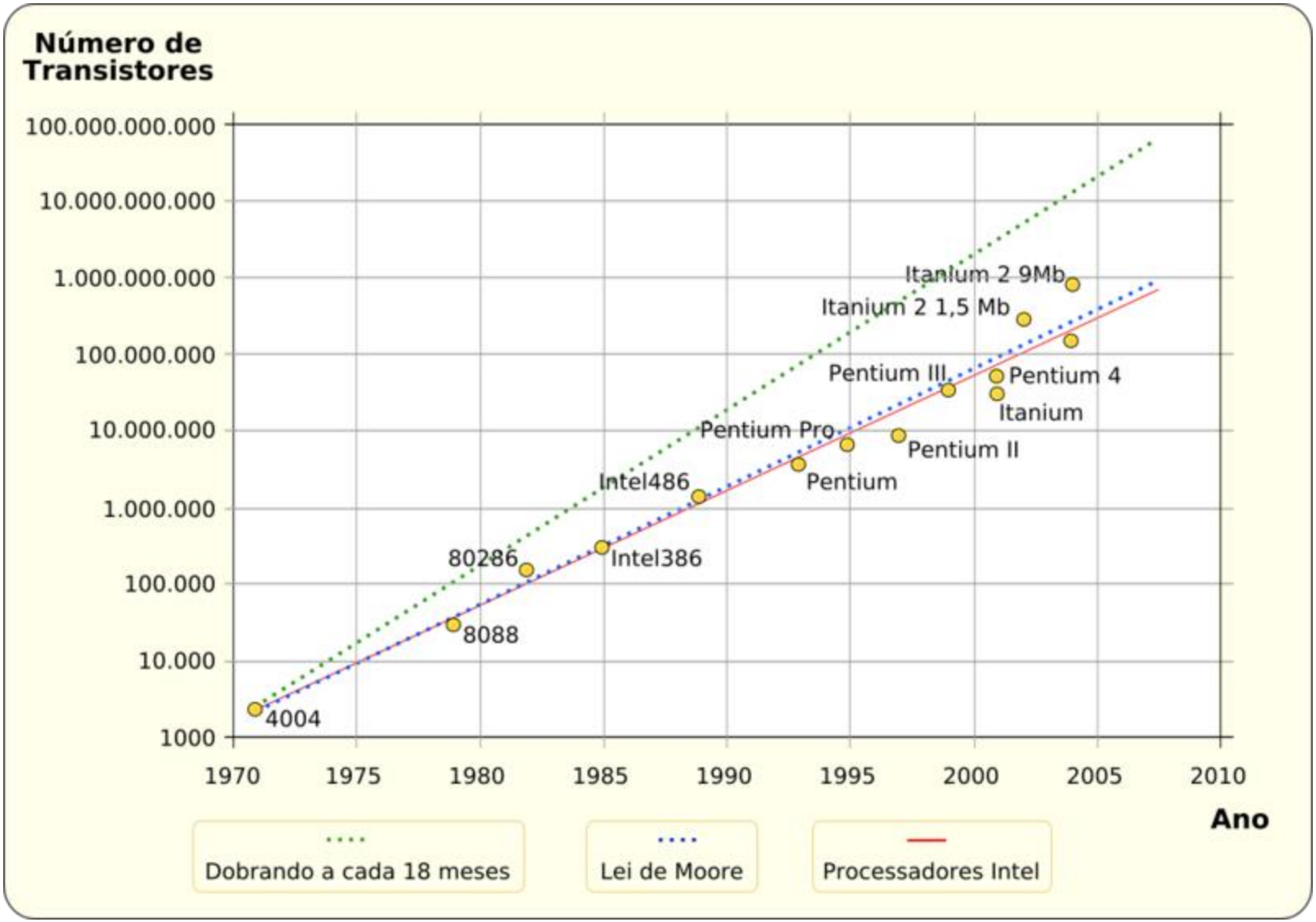
Integração das informações

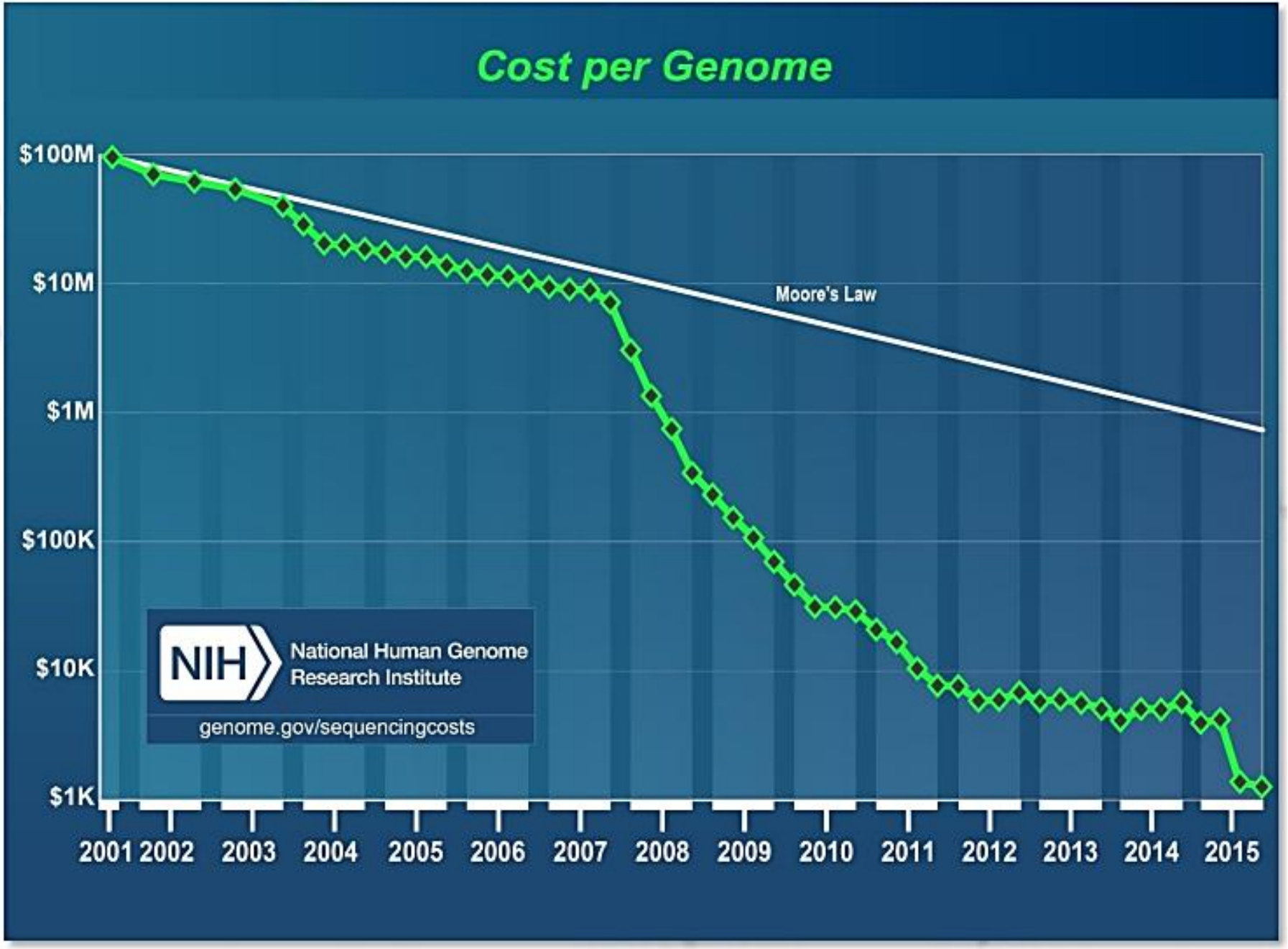
Estatística + Ferramentas Computacionais

Gerenciamento dos Dados

A Era do“Big Data”

A lei de Moore





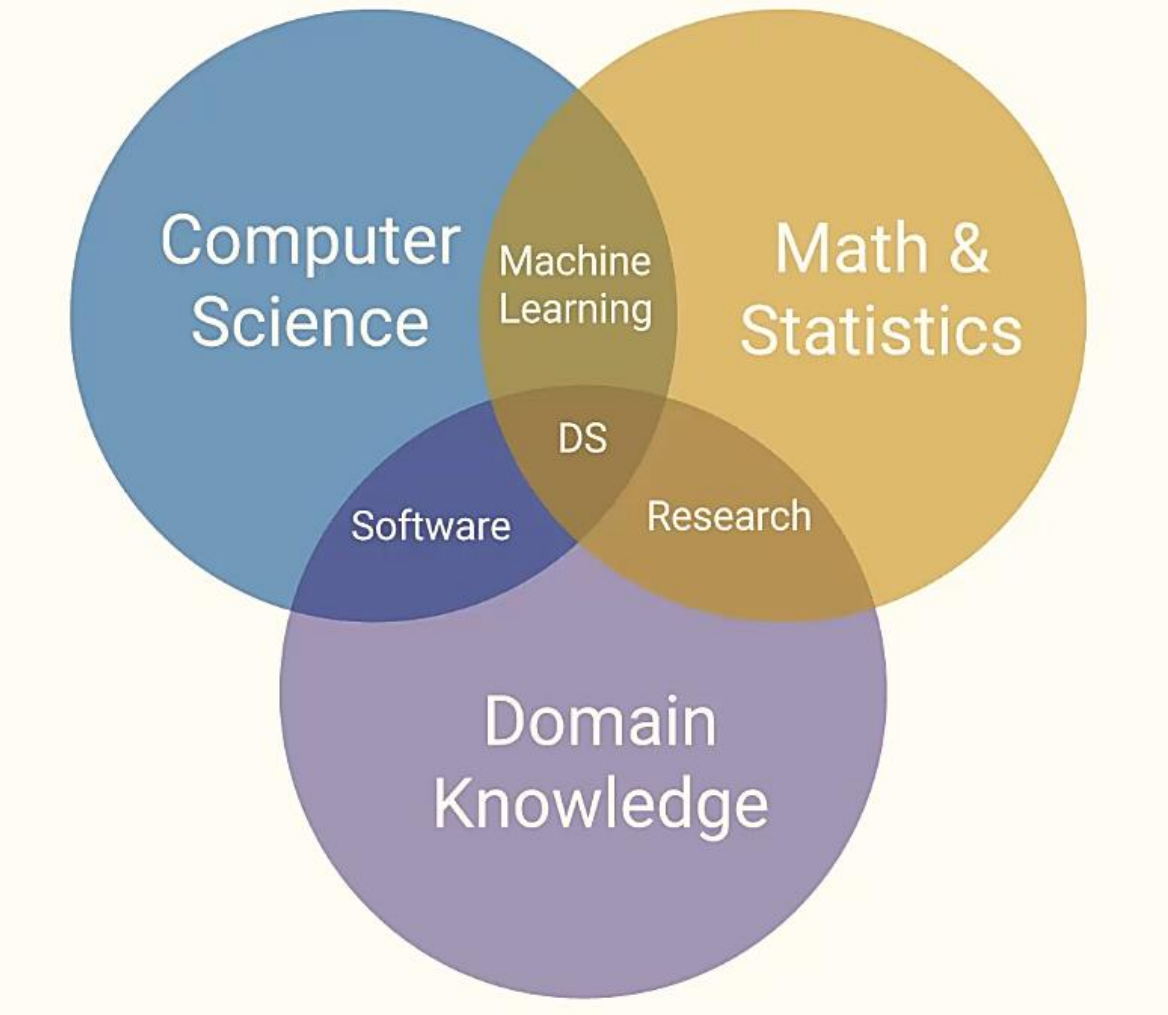
Fonte: Rosa, 2020

Data Science – Ciência dos dados

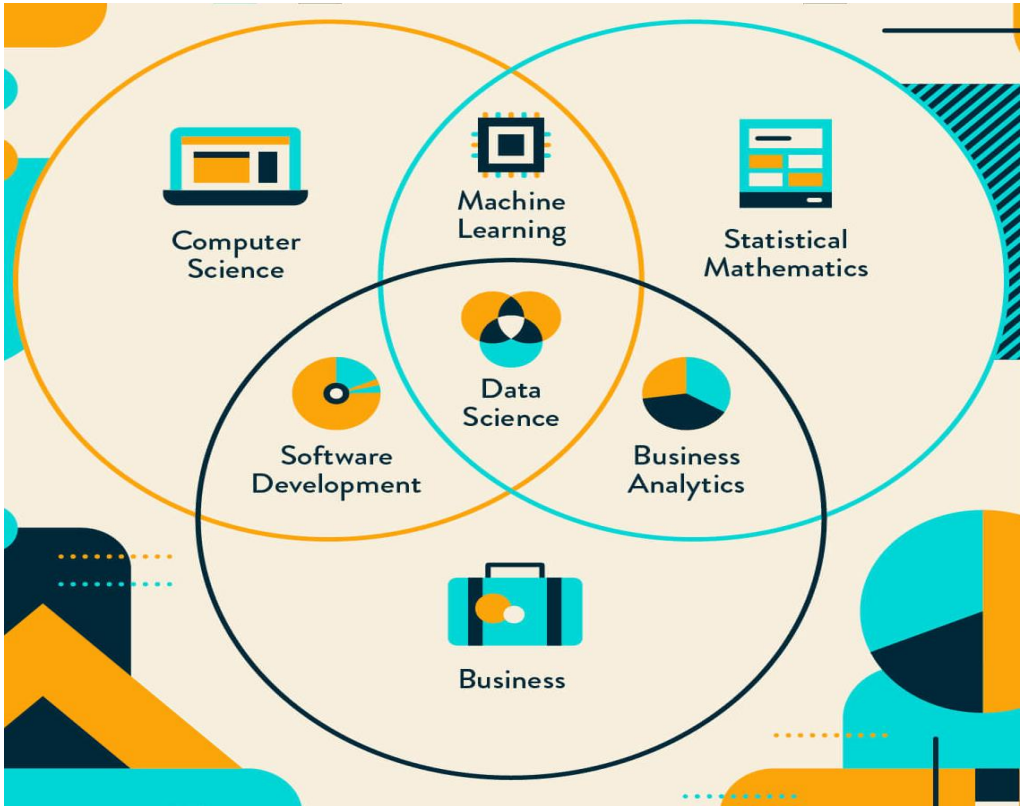
Por que do aumento na popularidade?

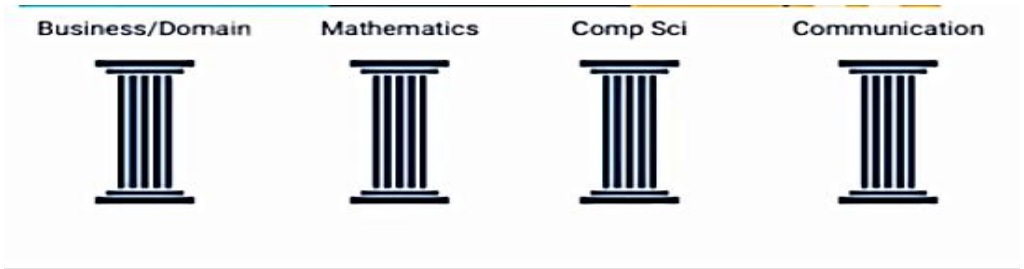
- Crescimento na disponibilidade de dados;
- Maior potência computacional (modelos não-lineares);
- Novas ferramentas de programação;
- Demanda de pessoas aptas para a ciência de dados;
- Grande aplicabilidade prática;

Data Science – Ciência dos dados



DATA SCIENCE





mathematics, statistics, computer science, domain knowledge and information science

Interdisciplinary Field

Uses scientific methods, processes, algorithms and systems

Extract knowledge and insights

from many structural and unstructured data.

Data science is a "concept to unify statistics, data analysis and their related methods"

understand and analyze actual phenomena

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory od Resources
Requirements, Assumptions and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion/Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Formal Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Asses Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results
Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions Decisions

Deployment

Plan Deployment
Deployment Plan

Plan Monitoring and Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience Documentation

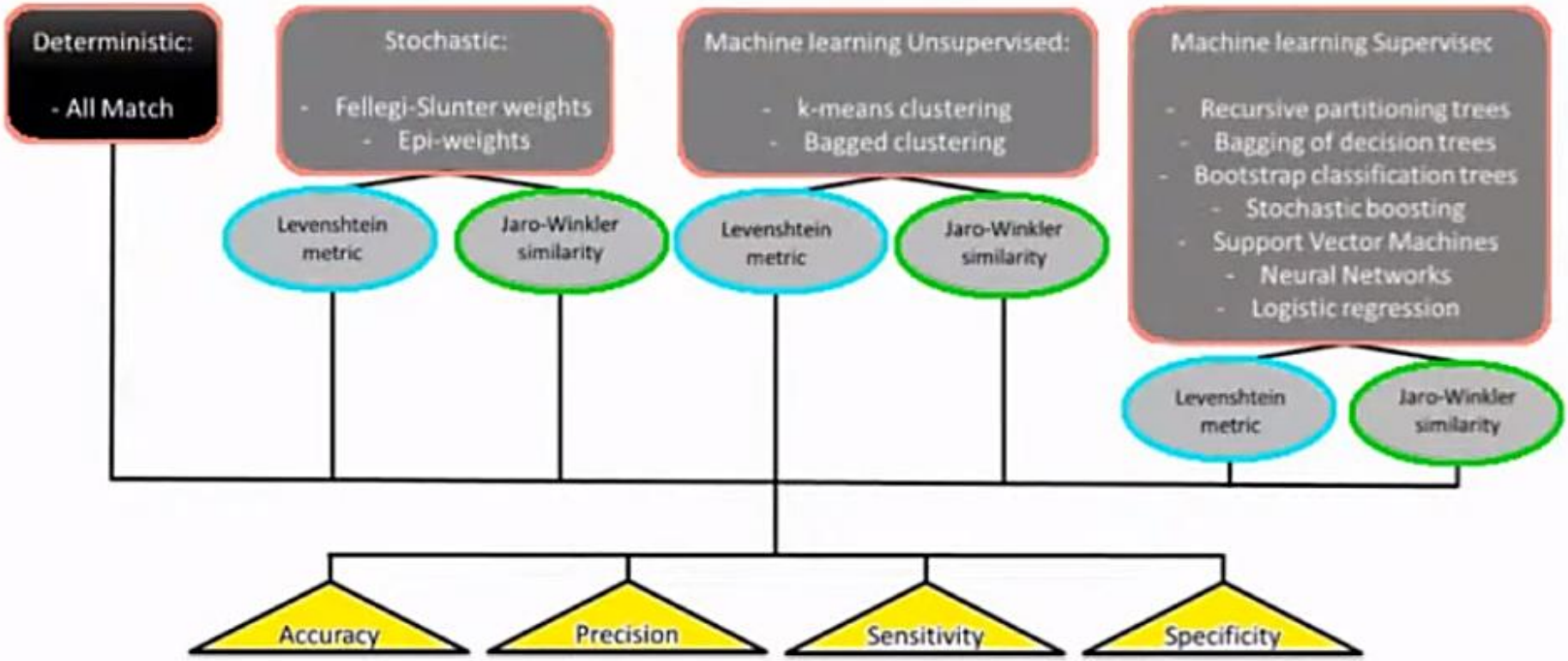
SUSTAINABILITY

Integração de Dados/centralizado/distribuído/icloud SQL Data-bases

	id	title	body	user_id
Edit Copy Delete	1	post one	hello from post one	1
Edit Copy Delete	2	this is post two	hello from post two	1
Edit Copy Delete	3	this is post three	hello from post three	2
Edit Copy Delete	4	this is post four	hello from post four	2
Edit Copy Delete	5	this is post five	hello from post five	2
Edit Copy Delete	6	this is post six	hello from post six	3
Edit Copy Delete	7	this is post seven	hello from post seven	3
Edit Copy Delete	8	this is post eight	hello from post eight	1

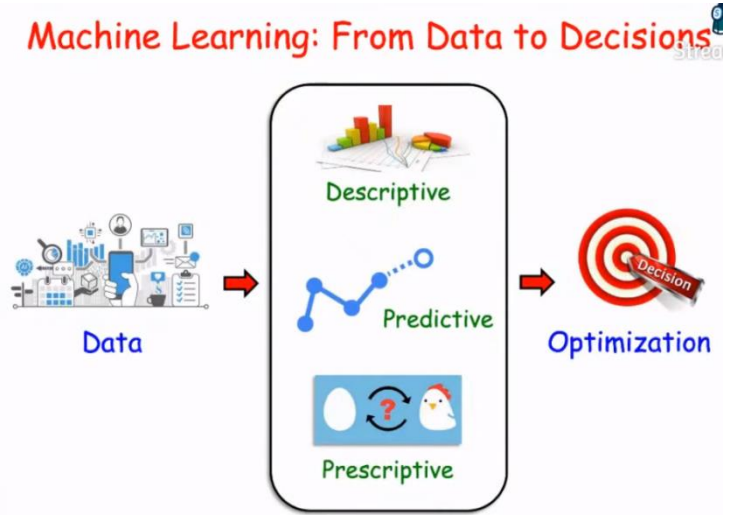
Data Integration

Comparison of Approaches for Farm Data Linkage
(Entity Matching)



Fonte: Rosa, 2020

Data Analytics vs Data Analysis



- **Data analysis é o procedimento que envolve investigar, limpar, transformar e treinar os dados com o objetivo de encontrar alguma informação útil, recomendar conclusões e ajudar na tomada de decisões;**
- **Data analytics envolve utilizar os dados, machine learning, análises estatísticas e ajustes de modelos computacionais a fim de obter uma melhor visão e melhores decisões a partir dos dados.**
- **“É o processo de transformar os dados em ações através de análises e discernimento no contexto de tomada de decisões e solução de problemas”.**

O que é Machine Learning?

- Machine Learning (aprendizado de máquina) é um método de análise de dados que automatiza a construção de um modelo analítico;
- Usa algoritmos que iterativamente aprendem dos dados;
- Machine learning permite que os computadores encontrem padrões nos dados sem explicitamente programar onde procurar esses padrões.

Gareth James

Home Bio Research Teaching CV Personal

*"Data is the sword of the 21st century, those who wield it well, the Samurai."**

Gareth James
Deputy Dean of the USC Marshall School of Business
E. Morgan Stanley Chair in Business Administration,
Professor of Data Sciences and Operations
Marshall School of Business
University of Southern California.

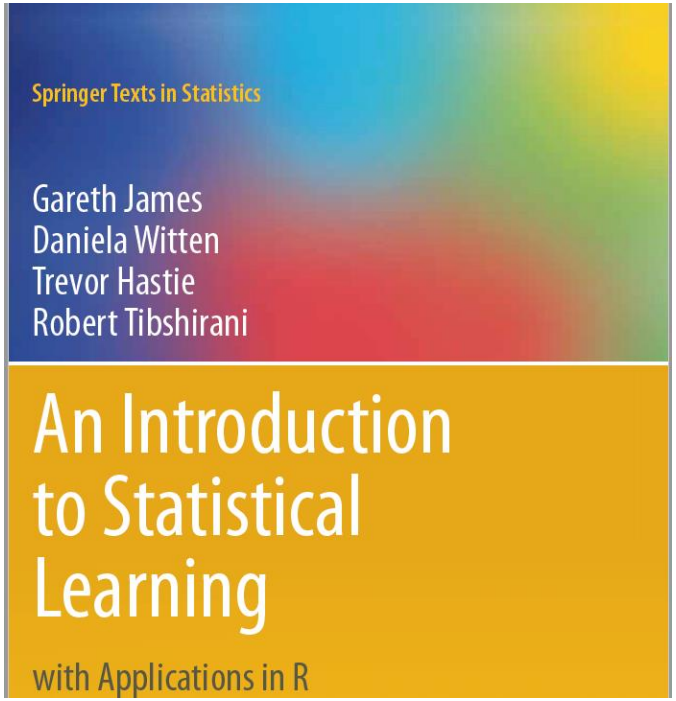
Education
BSc/BCom University of Auckland, New Zealand.
Ph.D. in Statistics, Stanford University, California.

Research Areas
Functional Data Analysis
High Dimensional Regression
Statistical Problems in Marketing

Contact Information
101 Bridge Hall
Data Sciences and Operations Department
University of Southern California.
Los Angeles, California 90089-0809
Phone: (213) 740 9696
email: gareth at usc dot edu

http://faculty.marshall.usc.edu/gareth-james/

Google: ISLR



STATISTICAL LEARNING = MACHINE LEARNING

"mathematical word"

- **Interesse apenas na teoria Matemática;**
- **Interesse apenas em esclarecer algo da teoria, porém aprofundando as aplicações no software R;**

- **Capítulo 1 e 2**

Competitions

Grow your data science skills by competing in our exciting competitions. Find help in the Documentation or learn about InClass competitions.

New to Kaggle? Start here!

Our Titanic Competition is a great first challenge to get started.

Titanic - Machine Learning from Disaster
Start here! Predict survival on the Titanic and get familiar with ML basics
Getting Started • Ongoing • 16999 Teams

Knowledge

All Competitions

Active | Completed | InClass

All Categories | Default Sort

Jane Street Market Prediction
Test your model against future real market data
Featured • 5 days to go • Code Competition • 4002 Teams

$100,000

HuBMAP - Hacking the Kidney
Identify glomeruli in human kidney tissue images
Research • a month to go • Code Competition • 1016 Teams

$60,000

RANZCR CLiP - Catheter and Line Position Challenge
Classify the presence and correct placement of tubes on chest x-rays to save lives
Featured • a month to go • Code Competition • 815 Teams

$50,000

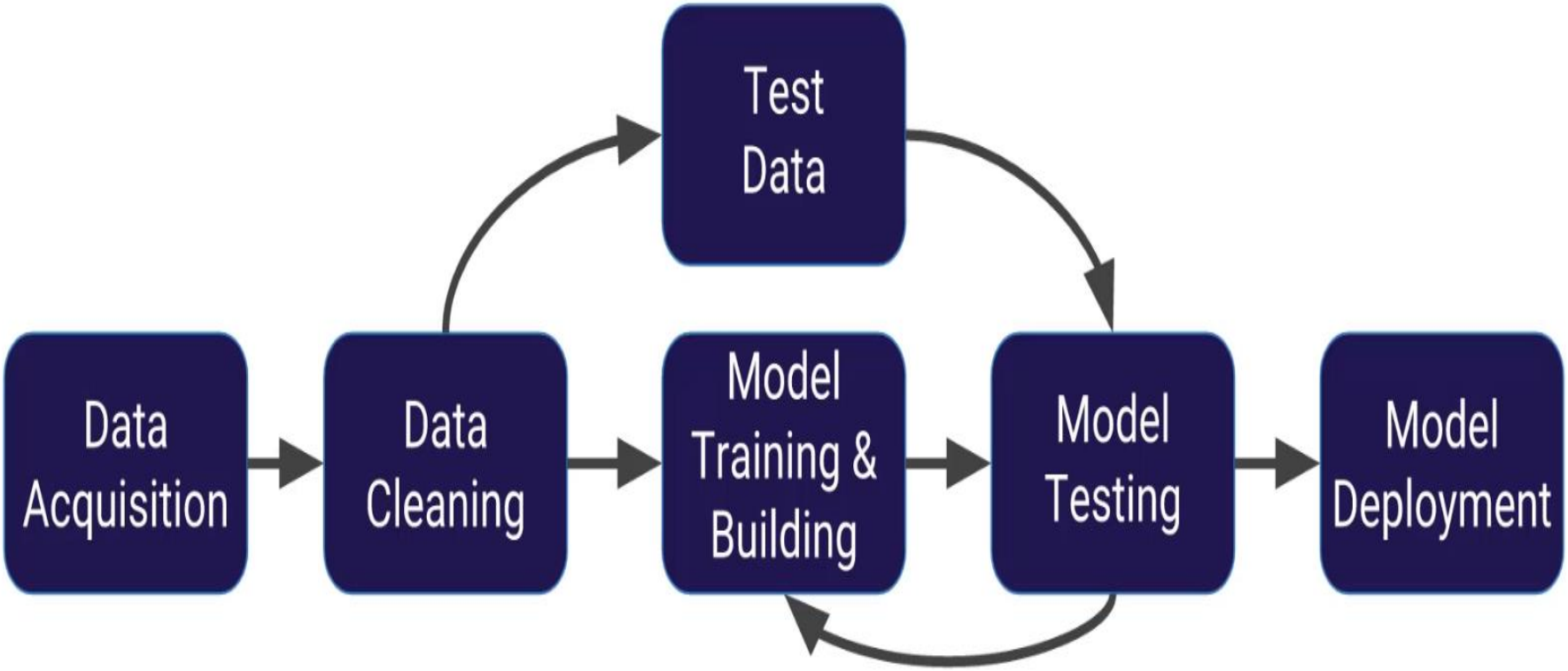
VinBigData Chest X-ray Abnormalities Detection
Automatically localize and classify thoracic abnormalities from chest radiographs

$50,000

Onde é utilizado?

Detecção de Fraudes	Segmentação de clientes
Resultado para pesquisa na Web	Reconhecimento de padrões de imagens
Propagandas na páginas da Web	Filtro de email e lixo eletrônico
Pontuação de créditos e próximas ofertas	Modelagem financeira
Predição das falhas de equipamentos	Genética e melhoramento Animal com dados de toda a cadeia
Modelos para nova precificação	
Detecção de intrusos na rede	

Processo geral



Training data

Formas de aprendizado/ Algoritmos

Supervised learning	Unsupervised Learning
Observações das variáveis e histórico dessas observações	Dados sem histórico/ se observam as variáveis de entrada.
Se deseja predizer um resultado ou inferir sobre as variáveis	O algoritmo precisa descobrir o que está sendo mostrado.
Para cada observação da variável existe uma variável resposta	Para cada observação da variável não existe uma variável resposta associada
Classificação, Regressão, GAMs (generalizes additive models),boosting	Agrupamento K-means; Decomposição singular, componentes principais, mapas de auto organização.
Ajuste sobre modelos de regressão linear	Não é possível o ajuste sobre modelos com regressão linear
Utilizado em aplicações onde dados históricos predizem prováveis eventos futuros.	Encontrar padrões de comportamento;
Exemplo: Prediz o preço de uma casa, baseado em diferentes características de diversas casas que possuem dados antigos; Dados de pesos de animais	Exemplo: Encontra principais atributos que separam segmentos de consumidores; Diversidade genética

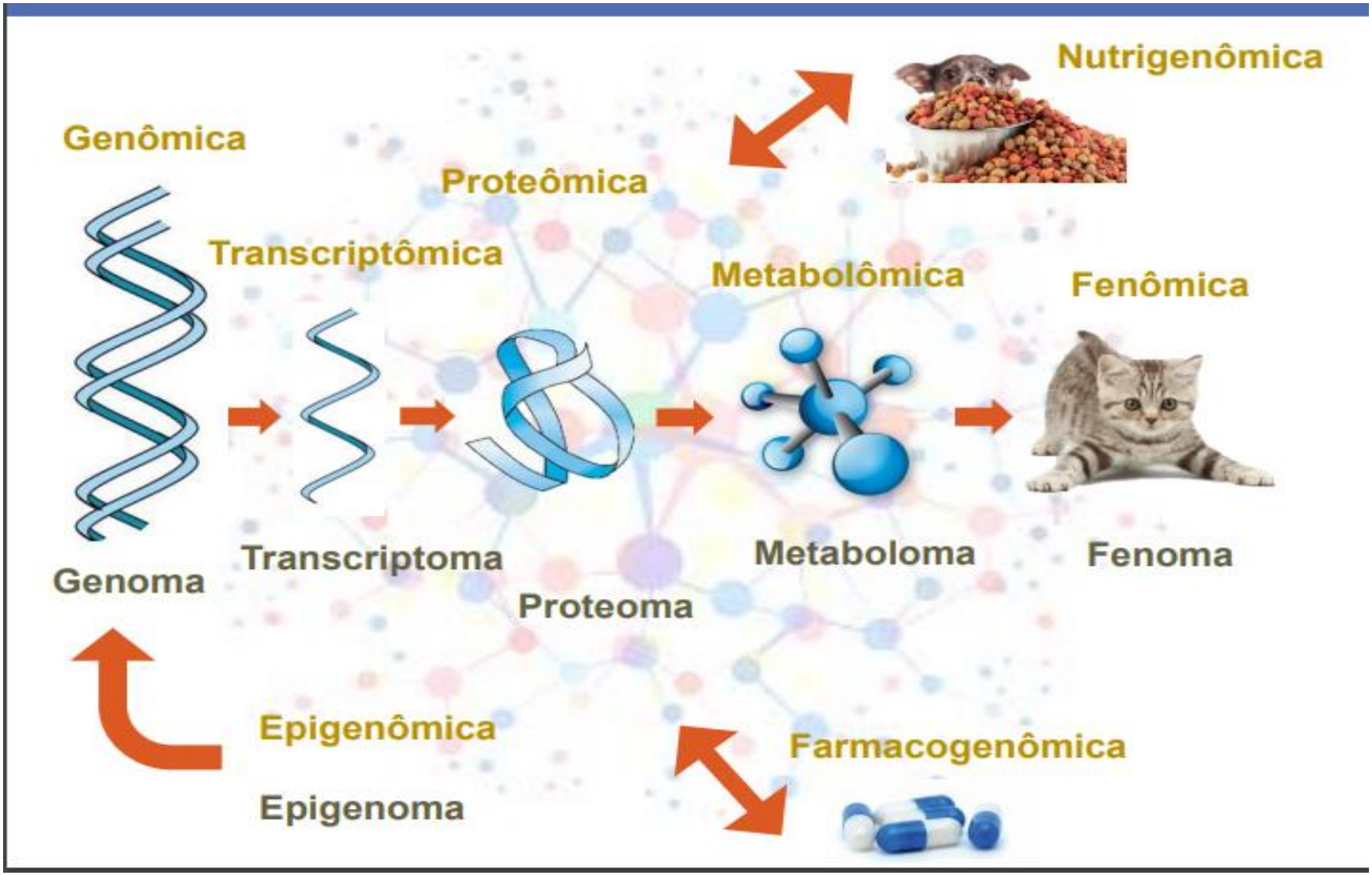
Reinforcement learning /Neural Network

- O algoritmo descobre, através de tentativa e erro, quais ações levam ao melhor resultado (prêmio).
- Robótica, jogos, navegação e caminhos de expressão gênica (ciências ômicas).
- Possui três componentes: o agente (o aprendiz ou tomador de decisões), o ambiente (tudo que interage com o agente) e ações (o que o agente pode ter).

ML no Melhoramento Genético Animal

- ML e MGA compartilham importantes objetivos tais como predição e muitas técnicas já são aplicadas em predição genômica;
- MGA utiliza de “Big Data” e métodos estatísticos – que entram dentro do escopo do ML, então o MGA é também machine learning; ou pelo menos uma parte da área de ML;

Tipo de informação – Bioinformação/Heterogeneidade



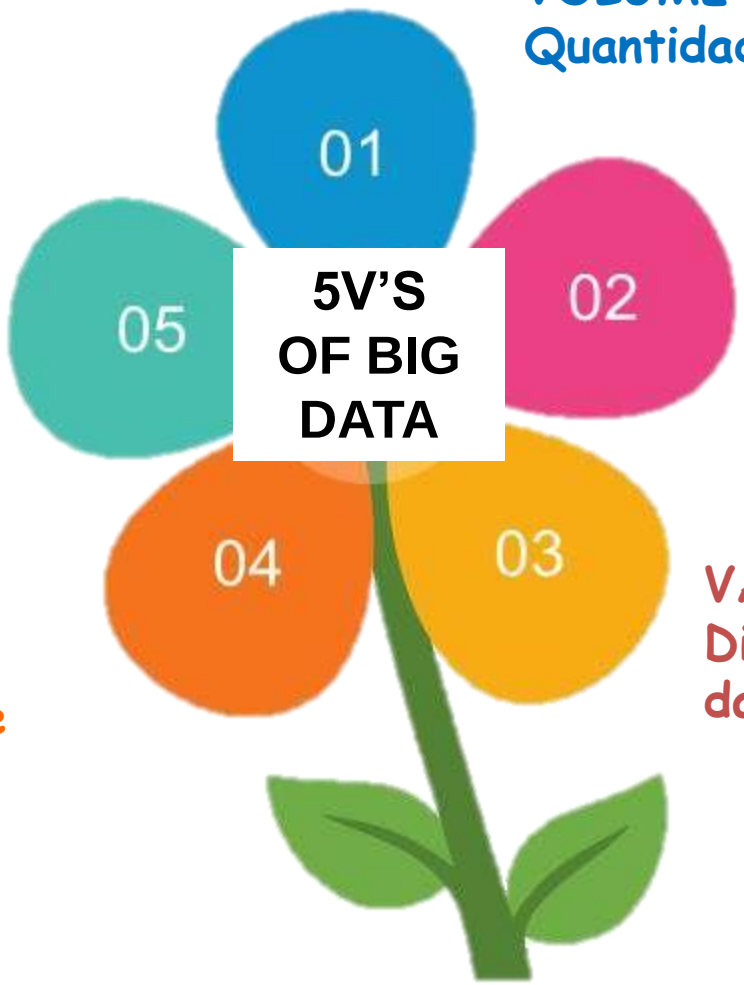
Mas o quanto é o “Big” Data?

CONCEITO RELATIVO

Classificação do que seria “aceitável” como base de dados

VOLUME –
Quantidade de dados

VALUE –
Ter uma grande quantidade de dados se torna inútil se não houver Valor neles

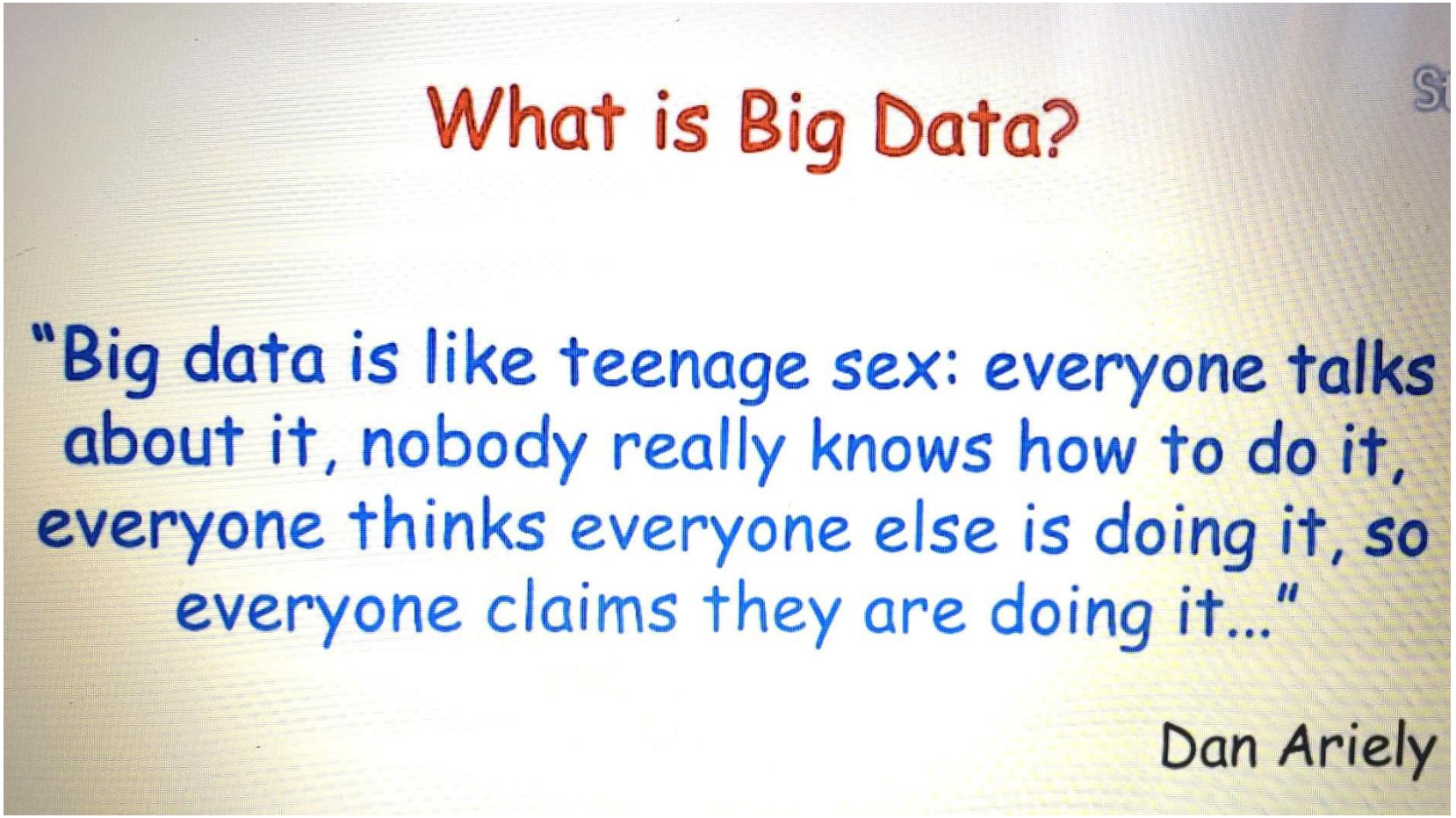


VELOCITY –
velocidade ao qual os dados são gerados

VERACITY –
Confiabilidade dos dados

VARIETY –
Diferentes tipos de dados

Temos uma definição “concreta”?



Fonte: Rosa, 2020

Inferência e Predição

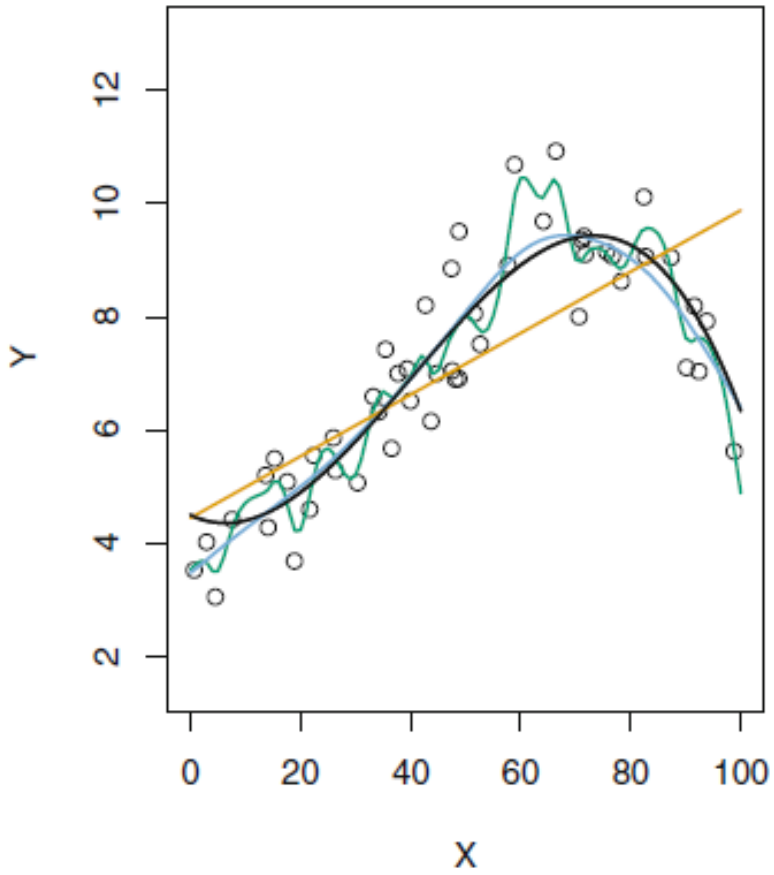
$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\
&= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\text{Var}(\epsilon)}_{\text{Irreducible}} ,
\end{aligned}
$$

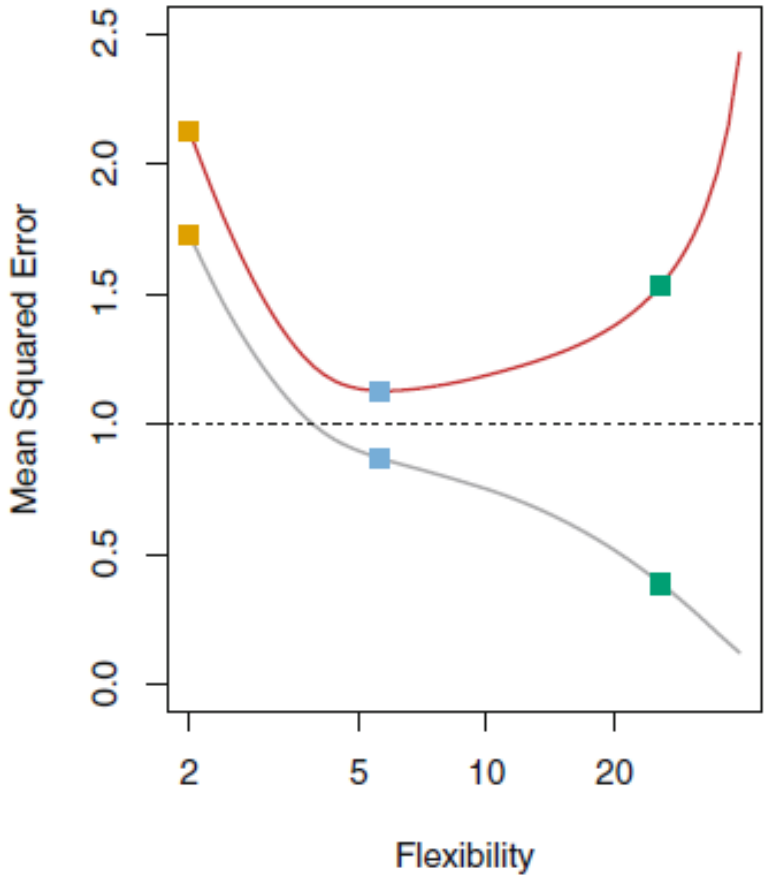
- Estimar **f** considerando a relação linear ou não-linear;
- Utilizar dados de treinamento “training data” para encontrar **f;**
- Encontrar um método de aprendizagem estatístico paramétrico ou não paramétrico

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$ **Training MSE**

Test MSE $$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon).$$

bias-variance trade-off





Critérios

- A escolha do adequado nível de flexibilização é um ponto crítico para o sucesso de qualquer método de aprendizado de máquina a ser escolhido.
- O Trade-off bias-variance e a resultante curva em forma de U no MSE, tornam a tarefa bastante complicada.

ML no Melhoramento Genético Animal

- ML pode ser utilizada para “aprender” a ponderar o efeito de cada SNP (prioris) e então utilizar esse aprendizado no método escolhido de predição genômica.
- Combinar dados estruturados com não - estruturados; (alto custo)
- Modelo considerado no MGA – modelo linear;
- Classification and regression trees, random forests, kernel-based methods e assim por diante.

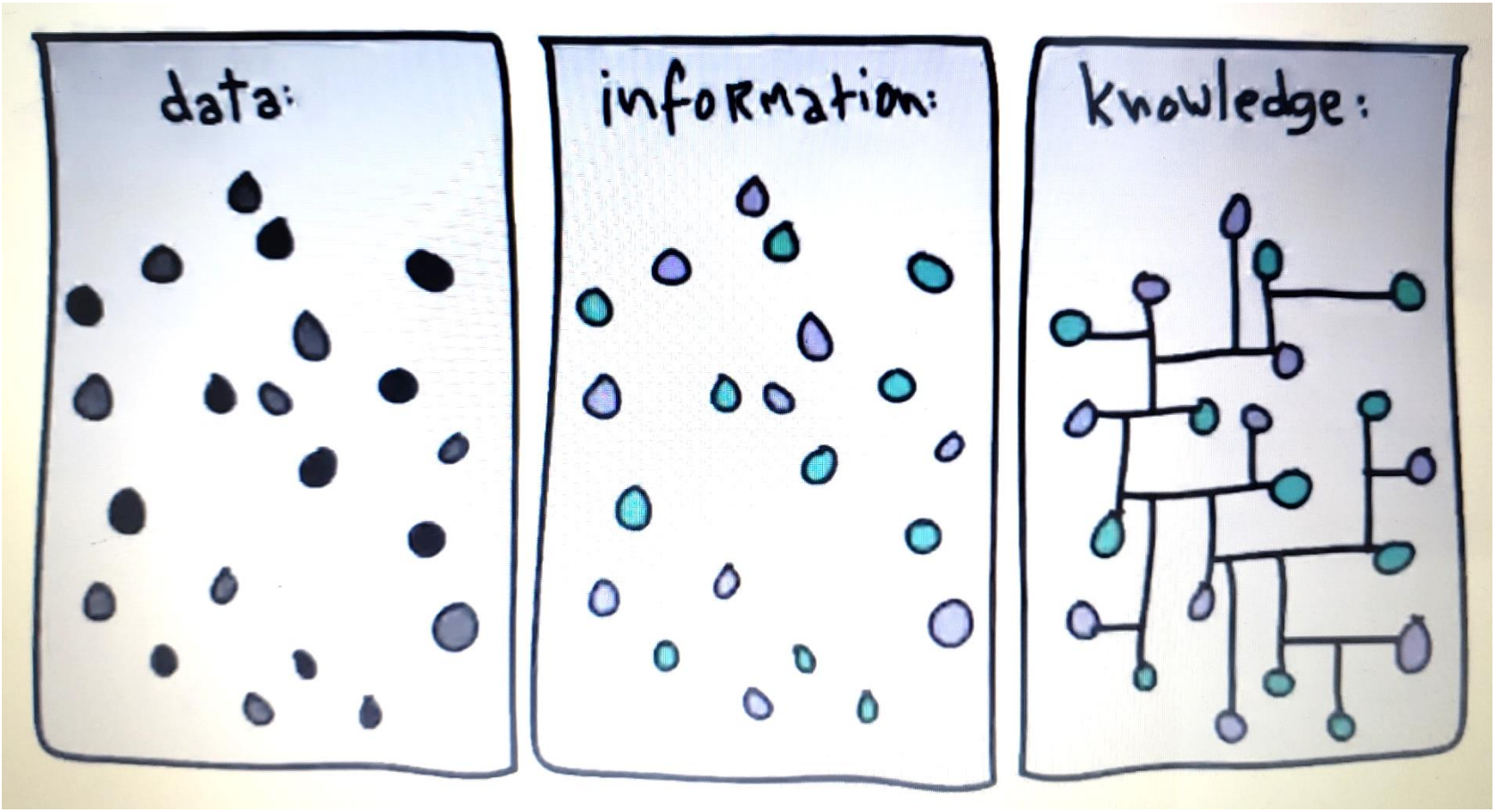
ML no Melhoramento Genético Animal

- MGA utiliza métodos condicionados ao modelo;
- ML procura encontrar o mais eficiente algoritmo e pouca importância a noção do modelo. (“model free”) – Hiperparametrização;
- Resultados difíceis de interpretar em ML;
- Entretanto, superestimando os sinais genéticos p > n;

ML no Melhoramento Genético Animal

- Pérez-Enciso M. **Animal Breeding learning from machine learning**. J Anim Breed Genet. 2017 Apr;134(2):85-86. doi: 10.1111/jbg.12263. PMID: 28297136;
- Long N, Gianola D, Rosa GJM, Weigel KA, Avendano S. **Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers**. Dev Biol (Basel). 2008;132:373-376. doi: 10.1159/000317279. PMID: 18817329.;
- Long N, Gianola D, Rosa GJ, Weigel KA, Avendaño S. **Comparison of classification methods for detecting associations between SNPs and chick mortality**. Genet Sel Evol. 2009 Jan 23;41(1):18. doi: 10.1186/1297-9686-41-18. PMID: 19284707; PMCID: PMC3225888.;
- SANT'ANNA, Isabela de Castro et al . **Subset selection of markers for the genome-enabled prediction of genetic values using radial basis function neural networks**. **Acta Sci., Agron.**, Maringá , v. 43, e46307, 2021 . Available from <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1807-86212021000102006&lng=en&nrm=iso>. access on 18 Feb. 2021. Epub Sep 11, 2020. http://dx.doi.org/10.4025/actasciagron.v43i1.46307.
- Cominotte, Alexandre & Araujo Fernandes, Arthur Francisco & Dorea, Joao & Rosa, Guilherme & Ladeira, MÃ¡rcio & Cleef, Eric & Pereira, G.L. & Baldassini, Welder & Machado Neto, Otavio. (2019). **Automated computer vision system to predict body weight and average daily gain in beef cattle during growing and finishing phases.** Livestock Science. 232. 103904. 10.1016/j.livsci.2019.103904.
- Vera Cardoso Ferreira Aiken, João Ricardo Rebouças Dórea, Juliano Sabella Acedo, Fernando Gonçalves de Sousa, Fábio Guerra Dias, Guilherme Jordão de Magalhães Rosa,. **Record linkage for farm-level data analytics: Comparison of deterministic, stochastic and machine learning methods,** Computers and Electronics in Agriculture, 2019, https://doi.org/10.1016/j.compag.2019.104857.

Então, o que podemos fazer?



Fonte: Rosa, 2020

Exemplo prático

- Dados.csv – Machine learning project.pdf

INFERÊNCIA CAUSAL