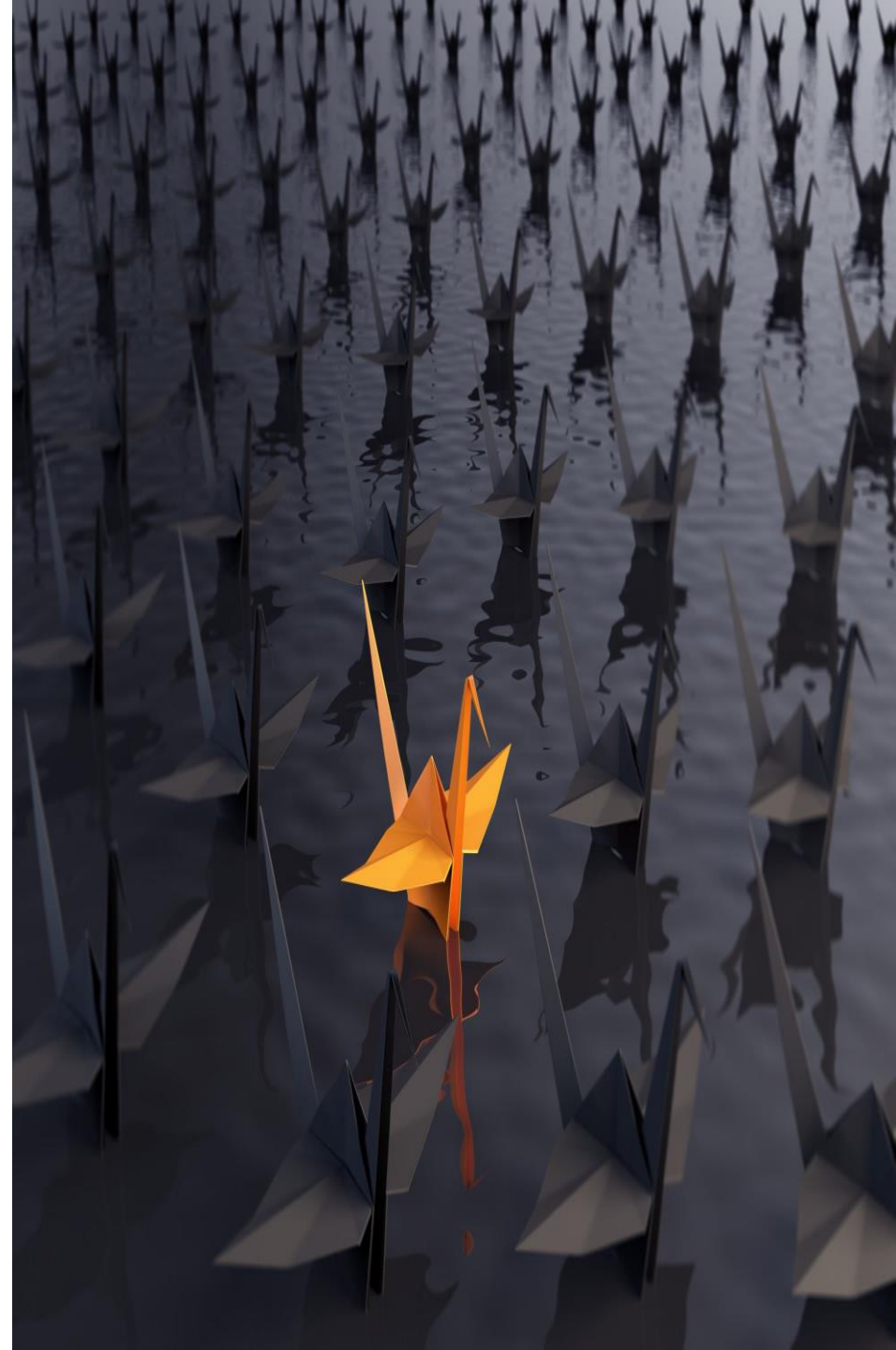


# Extração de Conhecimento de Bases de Dados

## *Knowledge Discovery in Databases (KDD)*

*Huei Diana Lee*

Inteligência Artificial  
CECE/UNIOESTE-FOZ



# Motivação

## **Passado**

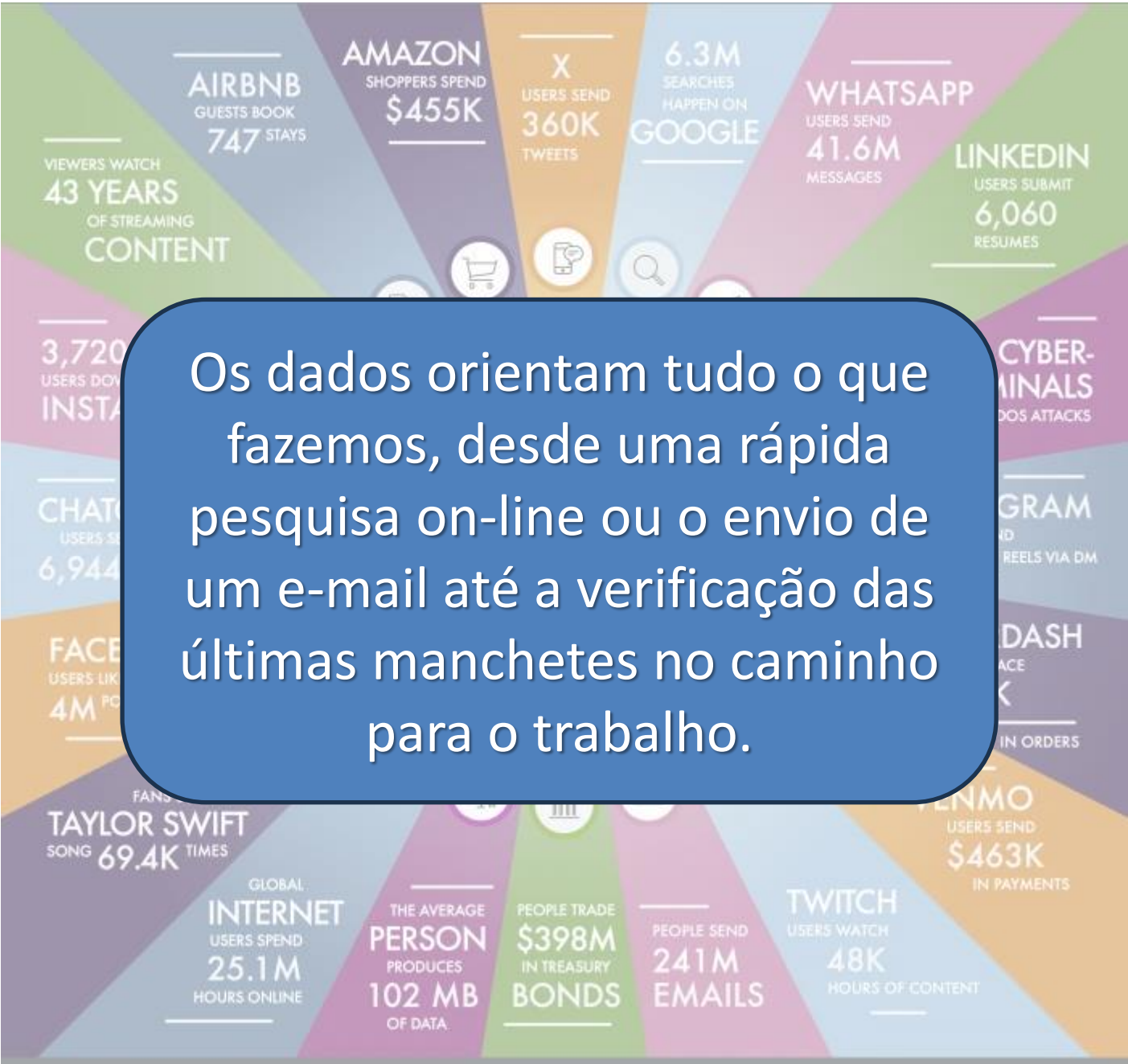
- Tecnologia limitada
- Armazenamento de pequenos volumes de dados (Mbytes)
- Consultas aos Dados
- Não existiam ferramentas para auxiliar a análise das informações obtidas

## **Presente/Futuro**

- Grandes avanços tecnológicos na área de TI
- Armazenamento de grandes volumes de dados (Tbytes, Pbyte...)
- Necessidade de conhecer e entender a BD
- O conhecimento extraído de uma BD deve ser usado para auxiliar as tomadas de decisões

# Etimologia

- Gigabyte ( $10^9$ ): Latim Gigas para Gigante
- Terabyte ( $10^{12}$ ): Grego Teras para Monstro
- Próximos prefixos:
  - Peta ( $10^{15} = 1000^5$ )
  - Exa ( $10^{18}$ )
  - Zetta ( $10^{21}$ )
  - Yotta ( $10^{24}$ )



Os dados orientam tudo o que fazemos, desde uma rápida pesquisa on-line ou o envio de um e-mail até a verificação das últimas manchetes no caminho para o trabalho.

# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>



# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>





# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>



# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>





# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>





# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>



# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>





# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>





# Data Never Sleeps 11.0

<https://www.domo.com/data-never-sleeps#>

# Data Never Sleeps 11.0

The world's internet population continues to grow significantly year-over-year. As of November 2023, the internet represents 5.2 billion people—approximately 64.6% of the global population. According to Statista, the total amount of data predicted to be created, captured, copied, and consumed globally in 2023 is 120 zettabytes, a number projected to grow to 181 zettabytes by 2025.

## Global Internet Population Growth (IN BILLIONS)



As data grows and evolves, businesses need to grow and evolve, too. Domo helps you harness the power of data so you can change as quickly as the world changes and make data-driven decisions that set you apart from the crowd. Let Domo help you make sense of all the clicks, swipes, and shares so you can see the big picture that a lot of small decisions make.

Learn more at [domo.com](https://domo.com)

SOURCES: EARTHWEB, DUSTIN STOUT, DEMANDSAGE, HOOTSUITE, BUSINESSOFAPPS, DOORDASH, SOCIALPILOT, X | TWITTER.COM, GITNIX, INVGATE, THINKIMPACT, SIFMA.ORG, STATISTA, PR NEWswire, NETSCOUT







# Data Never Sleeps 11.0

Domo has been keeping tabs on the world's data usage—in a minute—for over a decade now. What the numbers consistently show is that how we use data is always evolving—and that data isn't slowing down. We're also seeing some big changes. The rise of Artificial Intelligence (AI) is reshaping the way we communicate, work, and create. Digital payments continue to replace traditional transactions. Taylor Swift streams in countless headphones. And a rash of cybercrime grows alongside these digital experiences.

In Domo's 11th edition of Data Never Sleeps, we take the pulse of our digital age, where every click, swipe, and stream fuels an ever-expanding digital universe. These are not just numbers; they are the heartbeat of a world where data reigns supreme.



The world's internet population continues to grow significantly year-over-year. As of November 2023, the internet represents 5.2 billion people—approximately 64.6% of the global population. According to Statista, the total amount of data predicted to be created, captured, copied, and consumed globally in 2023 is 120 zettabytes, a number projected to grow to 181 zettabytes by 2025.

## Global Internet Population Growth (IN BILLIONS)



As data grows and evolves, businesses need to grow and evolve, too. Domo helps you harness the power of data so you can change as quickly as the world changes and make data-driven decisions that set you apart from the crowd. Let Domo help you make sense of all the clicks, swipes, and shares so you can see the big picture that a lot of small decisions make.

Learn more at [domo.com](https://domo.com)

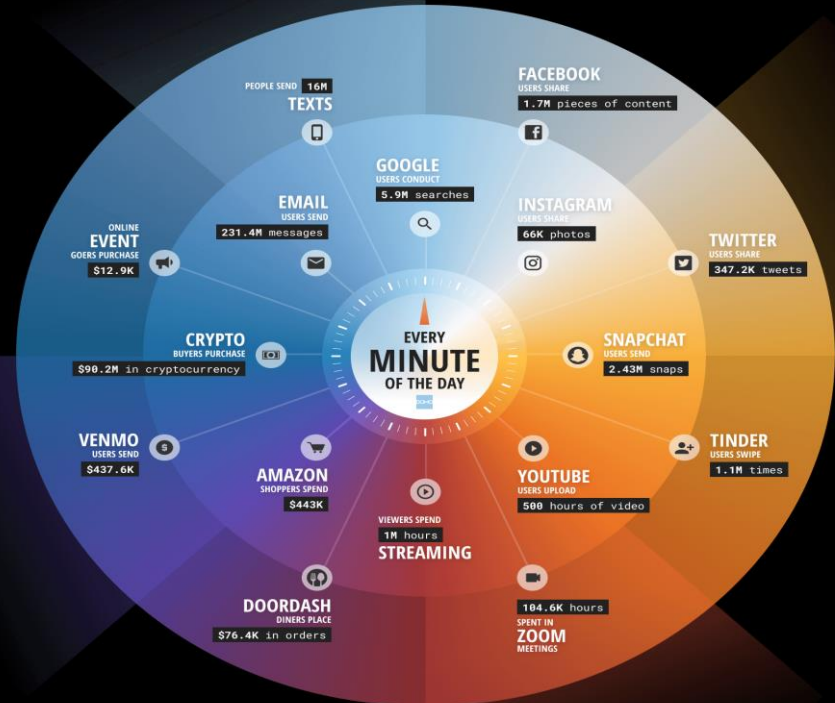
SOURCES: EARTHWEB, DUSTIN STOUT, DEMANISAGE, WOOTRICE, BUSINESSFAPPS, GOODRASH SOCIAL PILOT, X | TWITTER.COM, GITNIX, INVIGATE, THINKIMPACT, SIPMA.ORG, STATISTA, PR NEWSWIRE, NETSCOUT



# DATA NEVER SLEEPS 10.0

Over the last ten years, digital engagement through social media, streaming content, online purchasing, peer-to-peer payments and other activities has increased hundreds and even thousands of percentage points. While the world has faced a pandemic, economic ups and downs, and global unrest, there has been one constant in society:

our increasing use of new digital tools to support our personal and business needs, from connecting and communicating to conducting transactions and business. In this 10th annual "Data Never Sleeps" infographic, we share a glimpse at just how much data the internet produces each minute from some of this activity, marveling at the volume and variety of information that has been generated.



## DATA NEVER SLEEPS 1.0 VS. 10.0



## GLOBAL INTERNET POPULATION GROWTH (IN BILLIONS)



As of April 2022, the internet reaches 63% of the world's population, representing roughly 5 billion people. Of this total, 4.65 billion - over 93 percent - were social media users. According to Statista, the total amount of data predicted to be created, captured, copied and consumed globally in 2022 is 97 zettabytes, a number projected to grow to 181 zettabytes by 2025.

To succeed in an increasingly digital world where the volume of data created keeps accelerating, businesses need the right tools to put that data to work right where work gets done. Domo gives you the power to rapidly unlock value from all your data, regardless of where it lives, and drive actions across your organization that will improve business outcomes. Every click, swipe, share, or like tells a story, and Domo helps you do something powerful with it.

LEARN MORE AT [DOMO.COM](https://domo.com)

## SOURCES

Global Media Insights, Oberlo, Hootsuite, Earthweb, Matthew Woodward.co.uk, Web Tribunal, Deadline.com, Local IQ, Business of Apps, Query Sprout, Young and the Restless, Dating, Zillow, World, DoorDash, TechCrunch, Statista, Data Never Sleeps 1.0





# Gigantes, Monstros & “Leis”

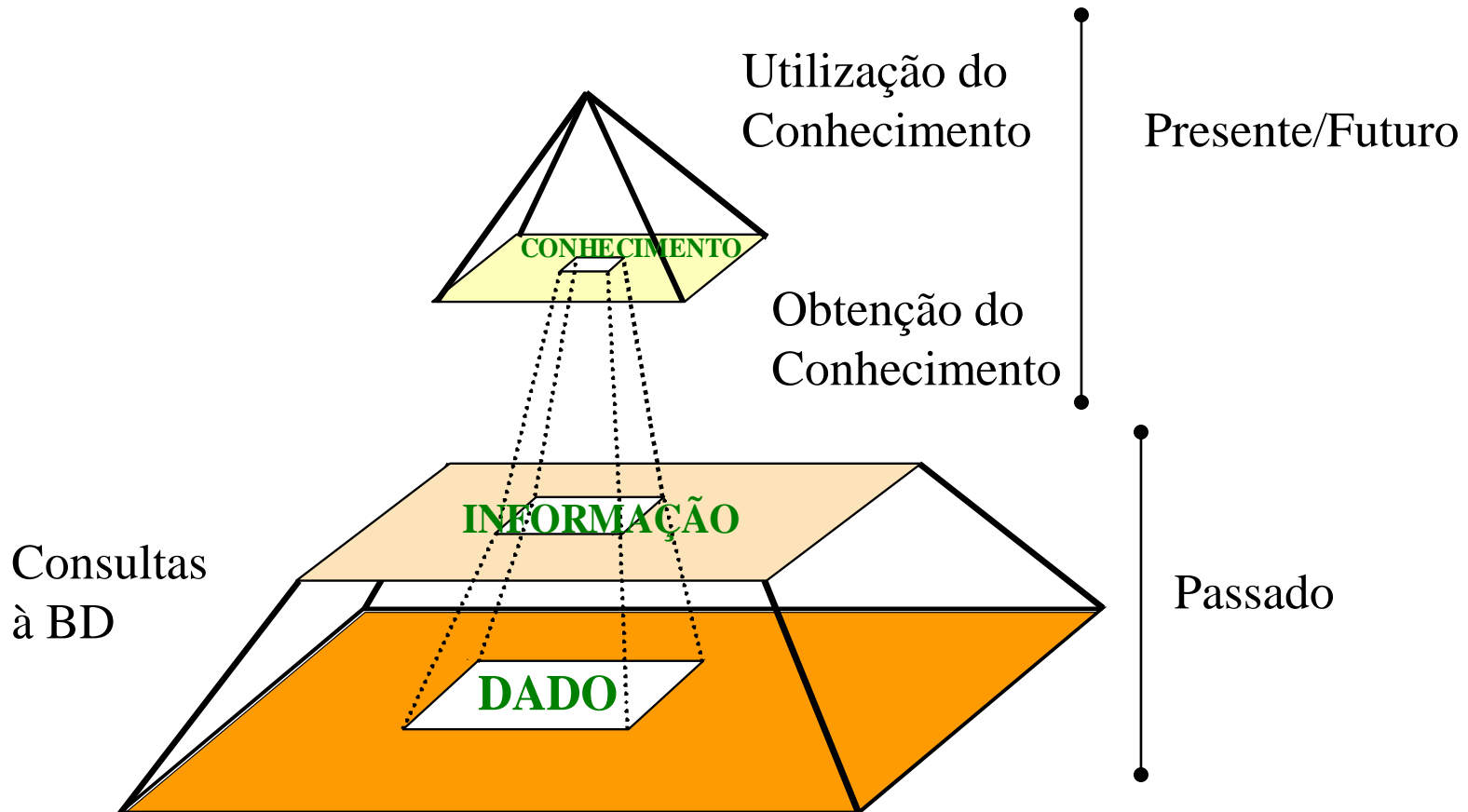
- Em 2015, 75% das empresas pesquisadas pretendiam investir em Big Data nos próximos 2 anos
- Objetivos:
  - Melhorar a experiência do cliente
  - Atingir mercados mais apropriados
  - Racionalizar processos existentes
  - Redução de custos
- Hype para valor

# O que é Big Data?



# Motivação

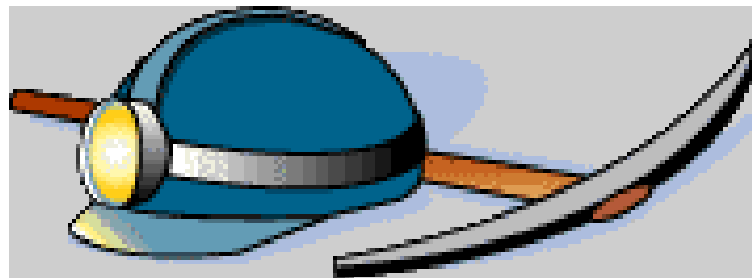
## Pirâmide do Conhecimento





# Introdução

O objetivo da extração de conhecimento é descobrir padrões, situações anômalas e ou interessantes, tendências e sequências nos dados



# Extração de Conhecimento de Base de Dados (KDD)

## KDD - Knowledge Discovery in Databases

- Pesquisadores norte-americanos
  - Criação de Métodos e Ferramentas
  - Auxiliar a Obtenção do Conhecimento
- KDD  $\neq$  Data Mining
- Processo de KDD

# Introdução



# Introdução

Qual produto de alta lucratividade venderia mais com a promoção de um item de baixa lucratividade, analisando os dados dos últimos dez anos?

Quais são os clientes potenciais para praticar fraudes?

Quais clientes gostariam de comprar o novo produto X?

Que genes são determinantes para o diagnóstico de um determinado tipo de doença?



# Exemplos de aplicações

# Big Data @ Walmart

## Walmart Big Data Facts and Figures



245 million customers visiting 10,900 stores and 10 active websites across the globe—Walmart is a name to reckon in the retail sector.

Walmart sees close to  
**300,000**  
social mentions every week.



Walmart sees close to  
**300,000**  
social mentions every week.



It has  
**2 million**  
associates and  
approximately half a  
million associates  
hired every year.



Walmart's employee numbers are more than some of the retailer's customer numbers.



Walmart takes in approximately **\$36 million** from across 4300 US stores every day.



Walmart collects  
**2.5 petabytes**  
of unstructured data from  
1 million customers every  
hour.



Walmart made a move from the experiential 10 node Hadoop cluster to a **250 node** Hadoop cluster in 2012.



Walmart has exhaustive customer data of close to 145 million Americans of which 60% of the data is of U.S. adults.



# Exemplos de aplicações



# Exemplos de aplicações



The analytics systems at Walmart analyse close to 100 million keywords on daily basis to optimize the bidding of each keyword.

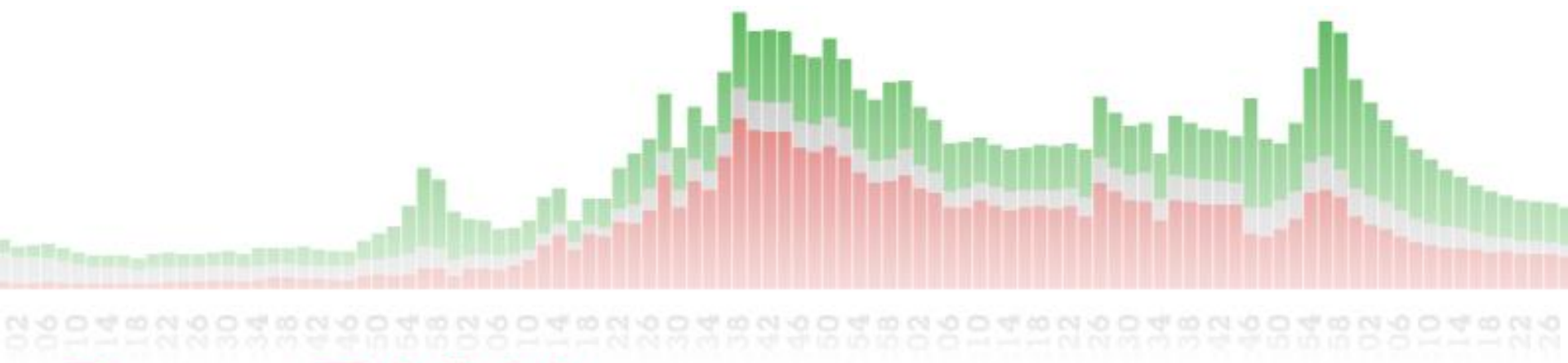
The analysis covers millions of products and 100's of millions customers from different sources.



Walmart observed a significant 10% to 15% increase in online sales for \$1 billion in incremental revenue.

Walmart Labs analyses every clickable action on Walmart.com –

- 1) What consumers buy in-store and online?
- 2) What is trending on Twitter?
- 3) Local events such as San Francisco giants winning the World Series?
- 4) How local weather deviations affect the buying patterns?



# DeepFAMA

Highly-scalable real-time social media sentiment analytics.





## DeepFAMA: High-Quality, High Volume Short Text Classifier

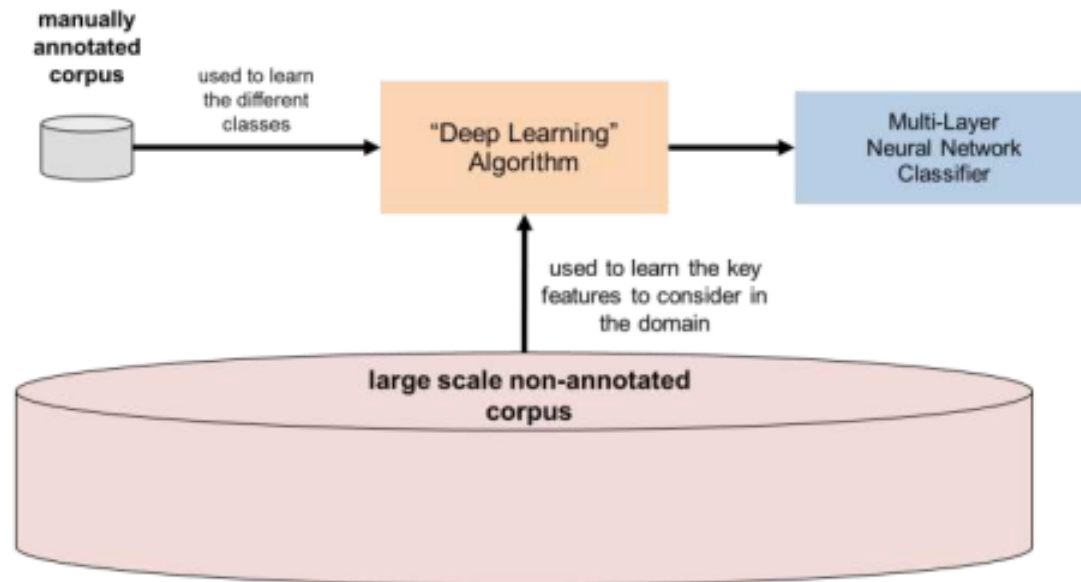
- *DeepFAMA* is a **short text classifier** developed by IBM Research – Brazil
- Applicable to conversational short texts (social media, SMS, call-center transcripts)
- Available in **English** and **Portuguese**;
- Implementation available for **high volume**, real time production scenarios (IBM Streams)
- **Human-level accuracy** achieved through new *Deep Learning* algorithm



**"FAMA"**

Greek goddess of gossip and rumor

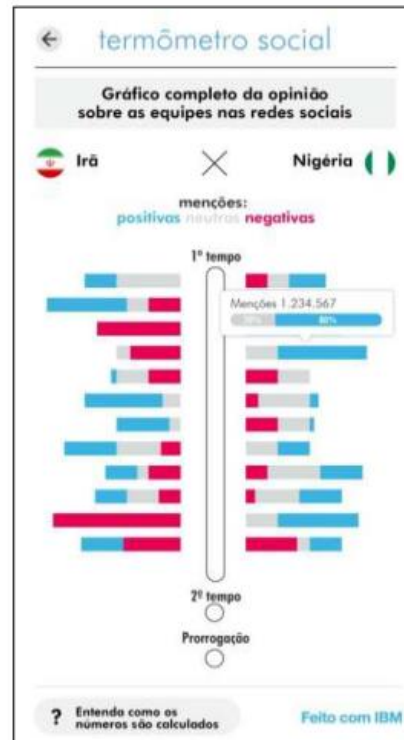
# New Deep Convolutional Neural Network (DCNN) Algorithm



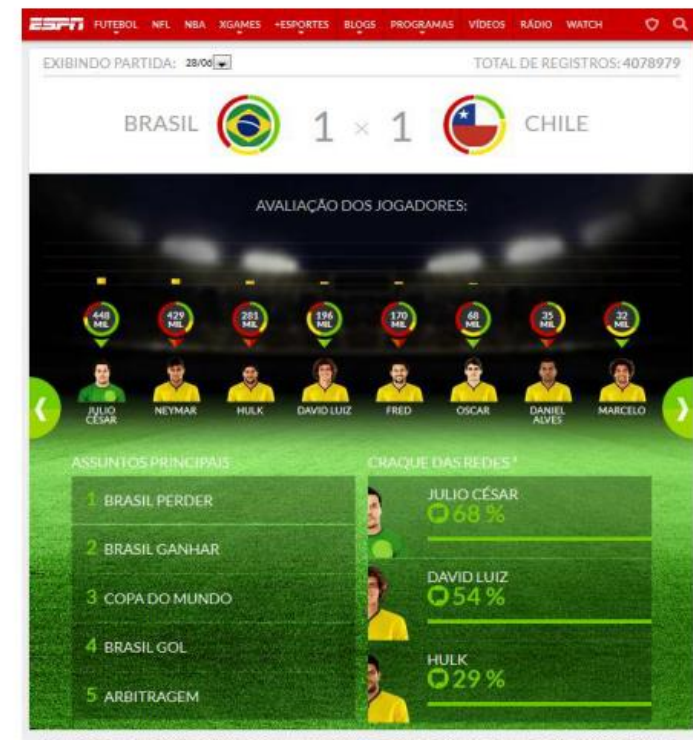
Cicero N. dos Santos and Maira Gatti. *Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts*. **Proceedings of COLING 2014**, pages 69–78, Dublin, Ireland, August 23-29 2014.

© 2015 IBM Corporation

# World Cup 2014 project with TV Globo, ESPN, and TV Band



**Globo 2<sup>nd</sup> screen app**  
1.4M downloads, 1.8M page views

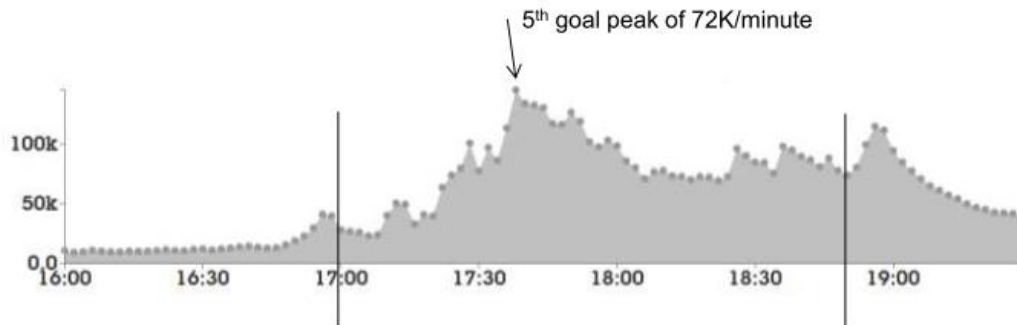


**ESPN Brazil**  
54.3K page views

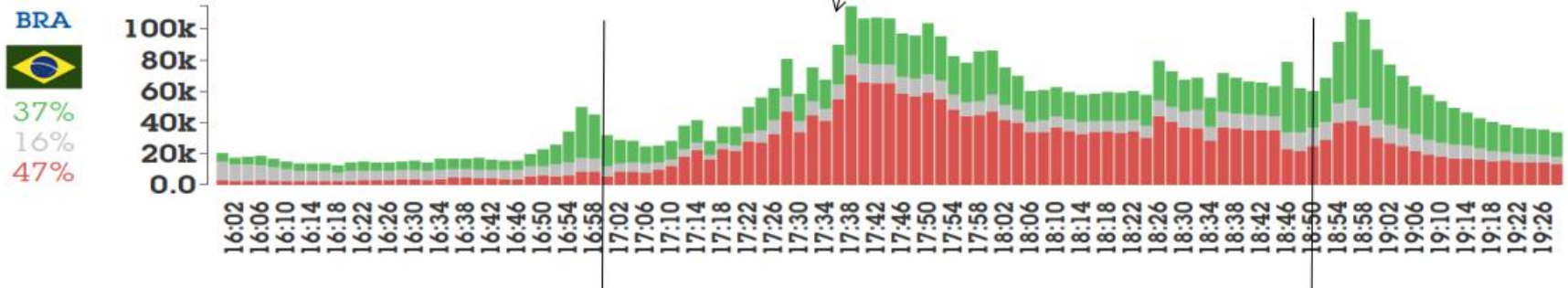
© 2015 IBM Corporation



# BRA 1x7 GER: Largest Event in SN History



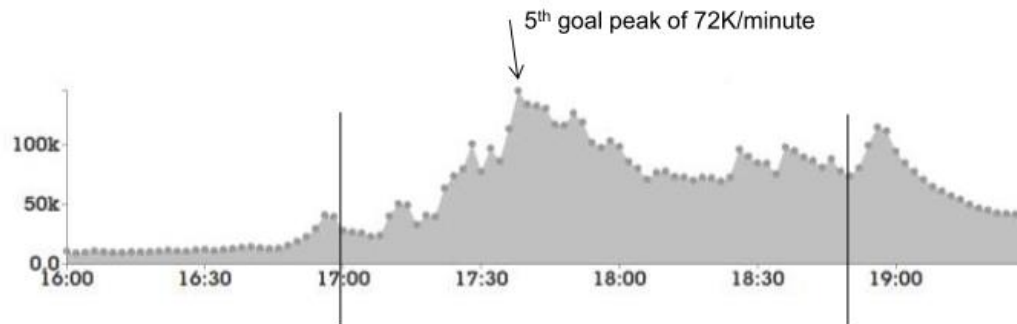
- globally 35.6M tweets (WR)
- 6.8M posts in Portuguese (19% of world)
- peak of 72K/minute
- 1.4M tweets after the game



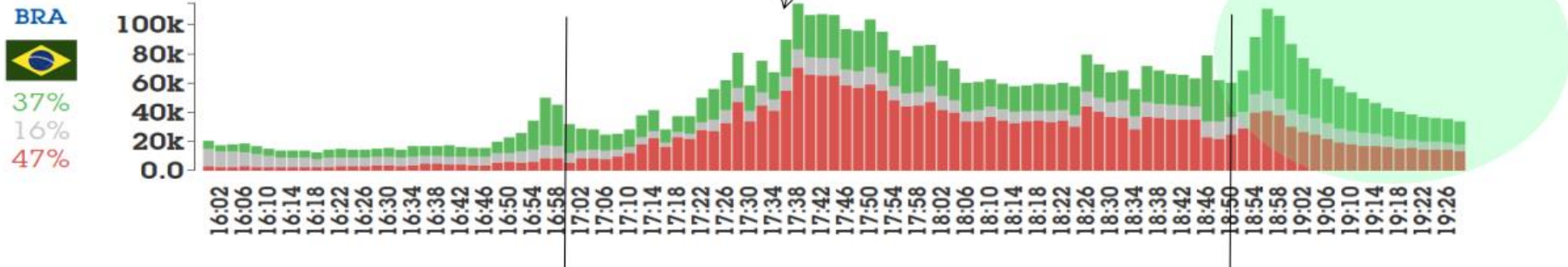
52

©Copyright 2014 IBM Co.

# BRA 1x7 GER: Largest Event in SN History



- globally 35.6M tweets (WR)
- 6.8M posts in Portuguese (19% of world)
- peak of 72K/minute
- 1.4M tweets after the game



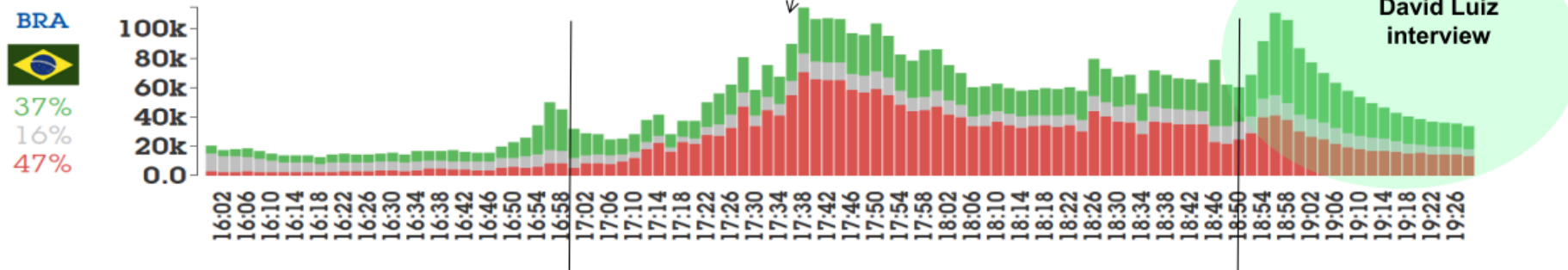
53

©Copyright 2014 IBM Co.

# BRA 1x7 GER: Largest Event in SN History



- globally 35.6M tweets (WR)
- 6.8M posts in Portuguese (19% of world)
- peak of 72K/minute
- 1.4M tweets after the game

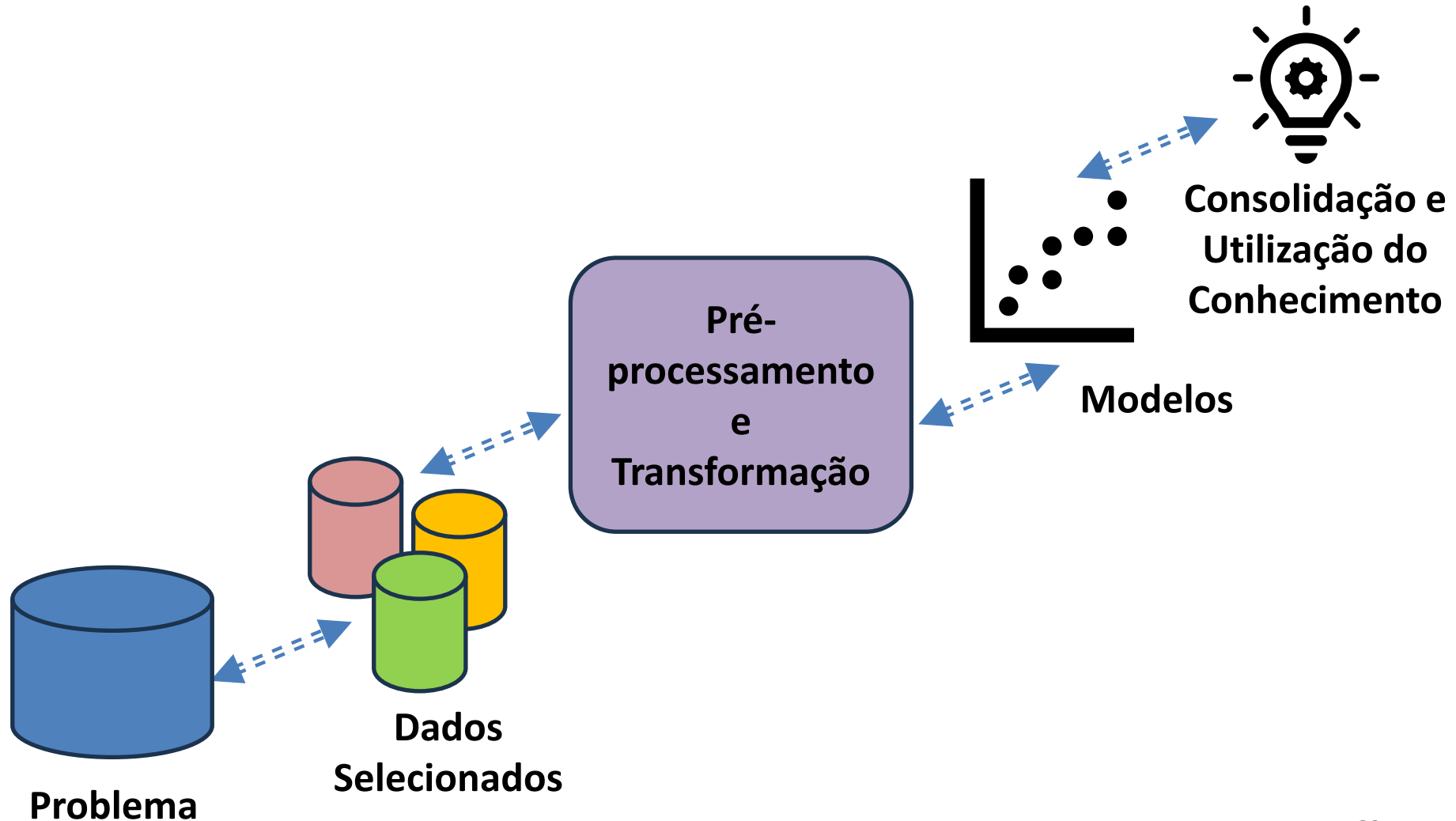


54

©Copyright 2014 IBM Corporation



# Processo de KDD

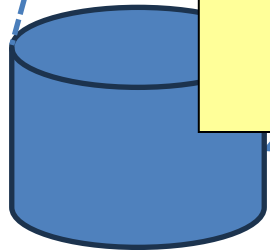


# Processo de KDD



Considera-se nessa etapa:

- Condições e metas do usuário final
- Estudo de viabilidades e custos da aplicação do processo
- Verificação do tipo e quantidade do conhecimento disponível antes de iniciar o processo de KDD
- Identificação dos gargalos do domínio
- Especificação do modo como o conhecimento extraído vai ser utilizado



**Problema**

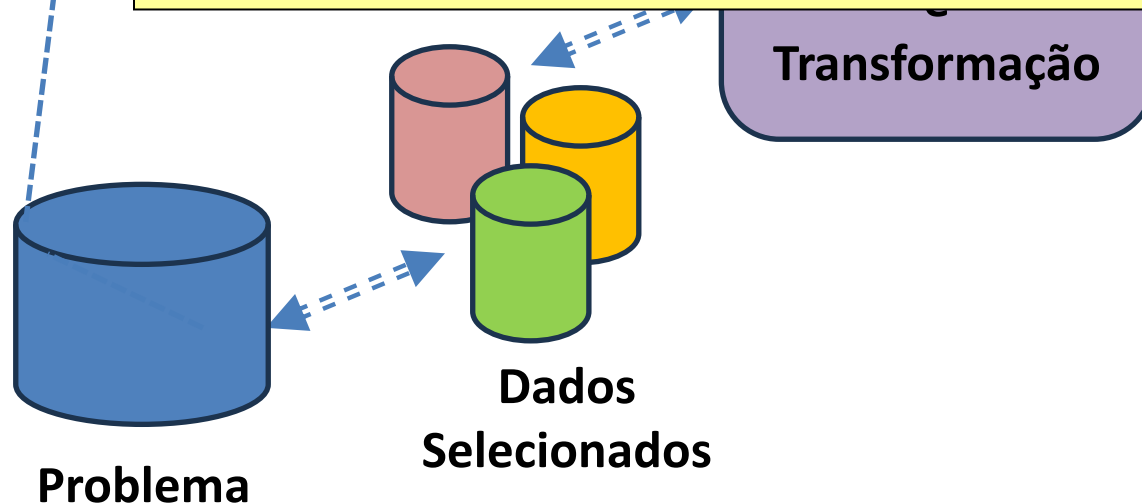
**Dados  
Selecionados**

# Processo de KDD

Alguns problemas da extração de conhecimento a partir de grandes dados:

- Limitação dos métodos de Data Mining quanto ao volume de dados
- Espaço de busca combinatoriamente explosivo
- Possibilidade de extração de padrões pouco significativos

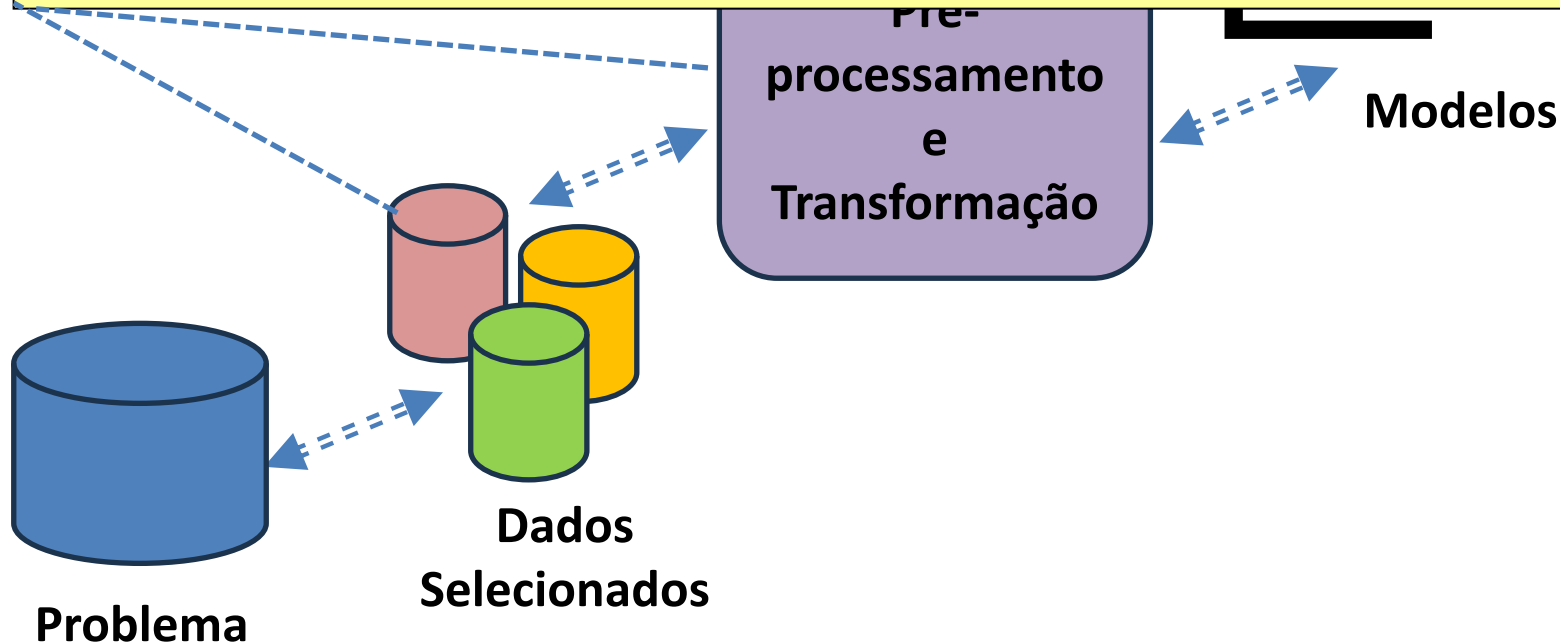
Esta etapa pode ser dividida em: seleção da amostra, e preparação e redução da amostra



# Processo de KDD

A seleção de uma amostra significativa considera os seguintes fatores:

- O tamanho da amostra
- Estratégias para obtenção da amostra
- Homogeneidade dos dados
- Dinâmica dos dados



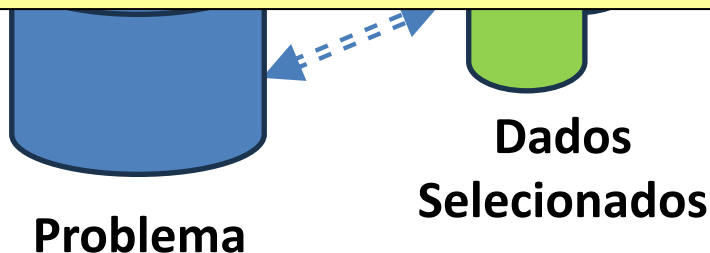


# Processo de KDD

Data Mining (DM) ou Mineração de Dados (MD) envolve a utilização de algoritmos para extração de padrões válidos, compreensíveis e potencialmente úteis nos dados.

Esses algoritmos consistem da combinação de três componentes:

- Modelo
  - Função do modelo
  - Representação do modelo
- Critério de preferência (*Bias*)
- Algoritmo de busca



# Processo de KDD

Pressupõe a verificação e a solução de potenciais conflitos com o conhecimento previamente extraído antes do processo iniciar.

O conhecimento extraído pode ser:

- Organizado pelo analista dentro de um novo modelo
- Utilizado para refinar um modelo existente ou
- Simplesmente documentado e informado ao usuário final



Consolidação e  
Utilização do  
Conhecimento

elos

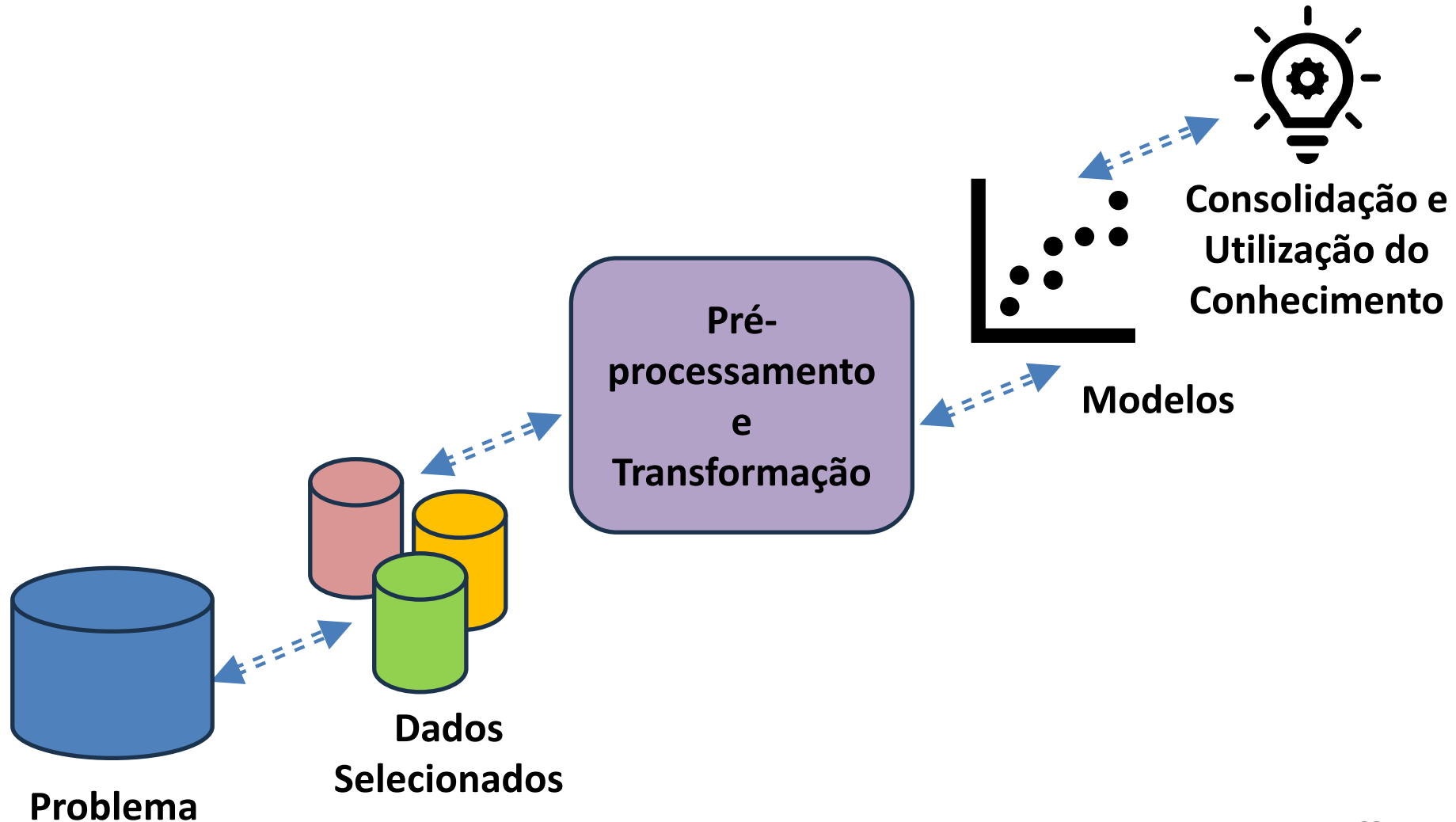


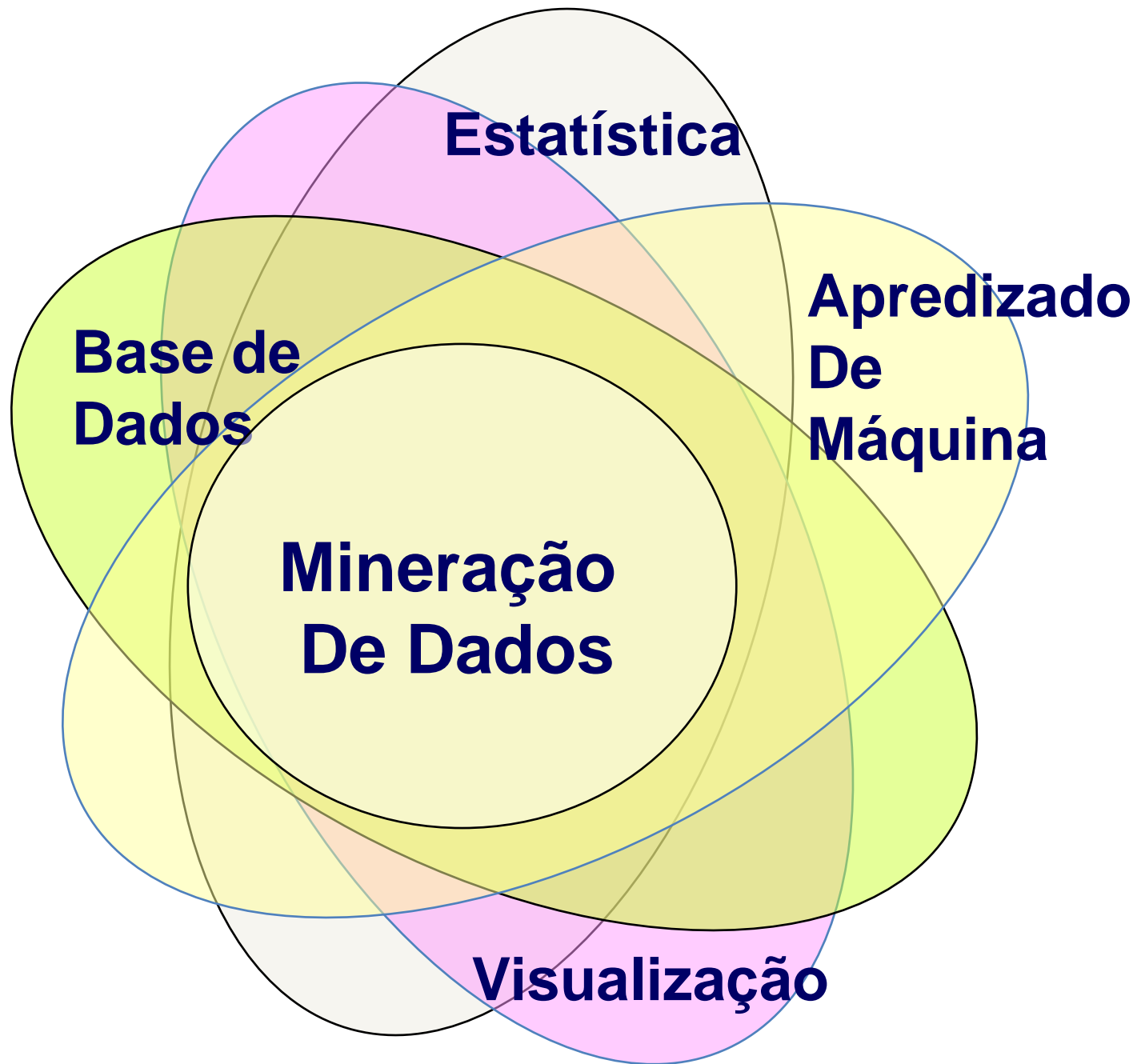
Problema



Dados  
Seleccionados

# Processo de KDD





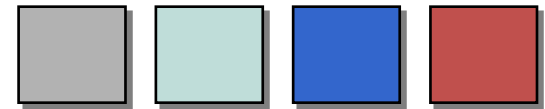


# Tarefas em Mineração de Dados

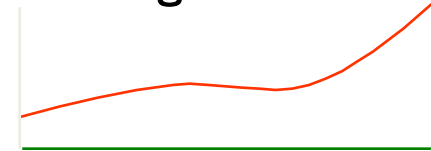
(focadas em Aprendizado de Máquina)

- Predição:
  - Classificação
  - Regressão
- Clustering
- Associação

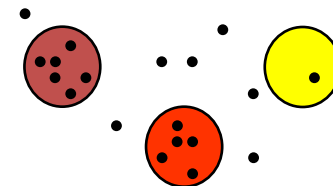
Classificação: Qual caixa?



Regressão



Clustering



Associação

	A	B	X	Δ
A				
B				
X				
Δ				

# Predição

- Estimativa ou prognóstico de um possível valor de um dado ausente
- Provável distribuição futura do valor baseado no conjunto histórico dos dados analisados
- **Exemplo:** potencial salário de um empregado pode ser previsto baseado na distribuição de salários de empregados com as mesmas características

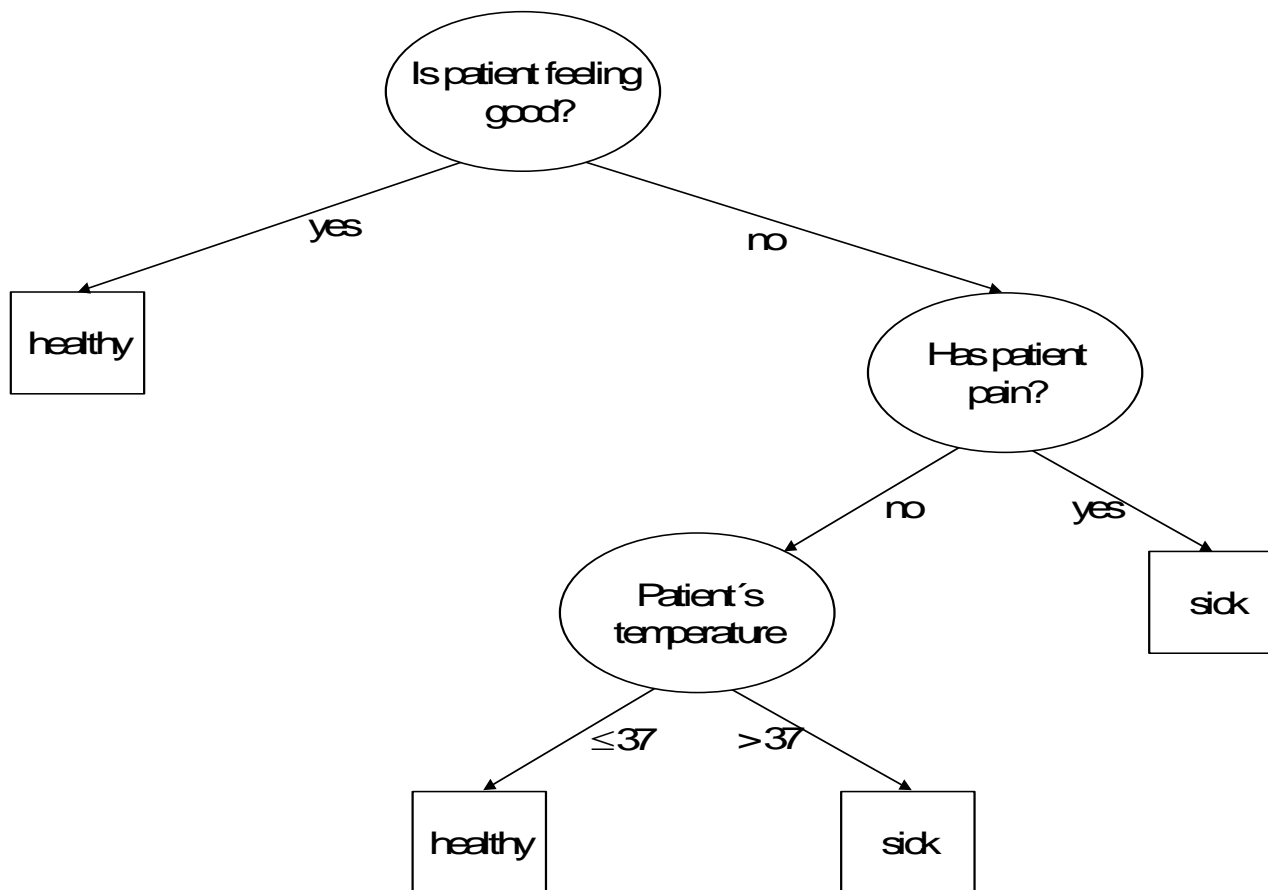


# Classificação

- Etiqueta, rótulo ou categoria de um dado em um conjunto de classes conhecidas
- Modelo de classificação é construído baseado nas características dos dados no conjunto treinado
- **Exemplo:** regras de classificação a respeito de doenças podem ser extraídas de um conjunto de casos conhecidos e usado para fazer um diagnóstico em novos pacientes baseado em seus sintomas



# Classificação





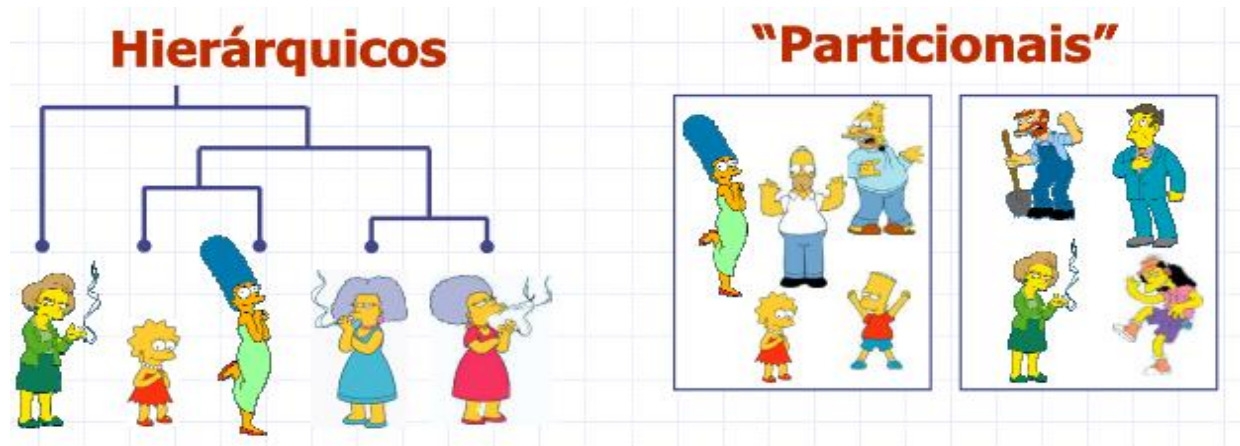
# Clustering

- **Categorização, segmentação ou agrupamento:** objetivo é agrupar objetos identificando grupos (clusters) baseadas em certos atributos
- **Critério de agrupamento:** maximizar as similaridades e minimizar as diferenças mediante algum critério
- **Exemplo:** um conjunto de novas doenças podem ser agrupadas em várias categorias baseadas nas similaridades de seus sintomas, e os sintomas comuns das doenças podem ser usados para descrever um grupo de doenças



# Clustering

- **Estratégias de Clustering:**
  - **Particionais:** construir várias partições e avaliá-las segundo algum critério
  - **Hierárquicos:** criar uma decomposição hierárquica do conjunto de objetos usando algum critério



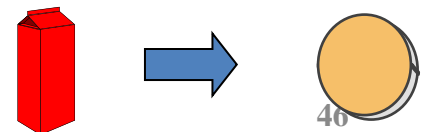
# Associação

- **Regras de associação:** tentam descobrir associações ou conexões entre objetos

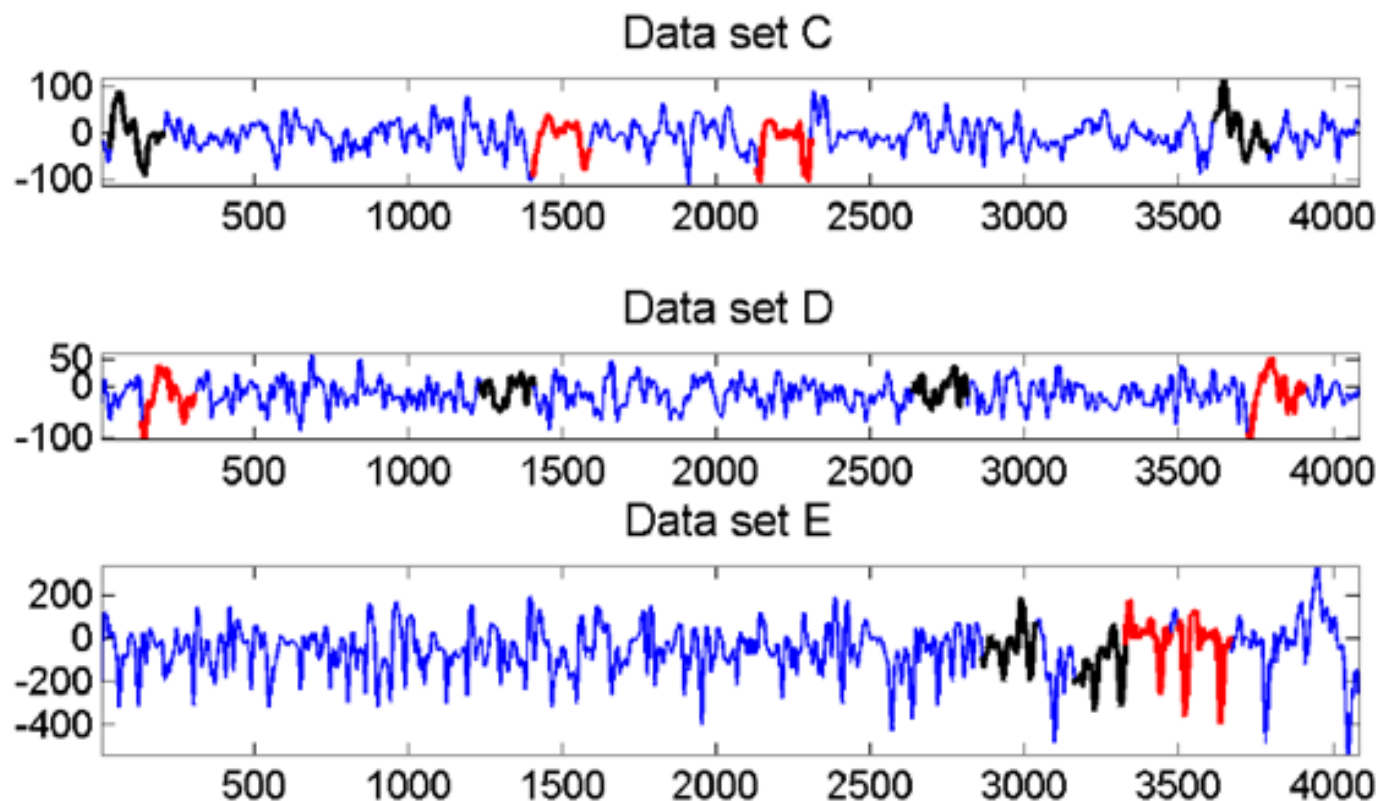
$$a_1 \wedge a_2 \wedge \dots \wedge a_n \rightarrow b_1 \wedge b_2 \wedge \dots \wedge b_n$$

significa que os objetos  $b_1 \wedge b_2 \wedge \dots \wedge b_n$  tendem a aparecer com os objetos  $a_1 \wedge a_2 \wedge \dots \wedge a_n$  dentro de um conjunto de dados

- **Exemplo:** pode-se descobrir que um conjunto de sintomas acontece com frequência junto a um outro conjunto de sintomas, e então, estudar os motivos dessa associação



# Evolução





# Ferramentas

- Várias ferramentas comerciais:
  - Relativamente caras
  - Maioria não apresenta suporte para todas as etapas de KDD
  - Aproveitando a “onda data mining”
- Centros de pesquisas e empresas desenvolvem ferramentas de domínio público

# Ferramentas

- Ferramentas Comerciais:
  - MineSet™ - Silicon Graphics
  - Enterprise Miner™ - SAS Institute
  - Intelligent Miner™ - IBM
  - Orange
  - Pentaho
- Ferramentas de Domínio Público:
  - Pentaho
  - Orange
  - WEKA - Univ. de Waikato na Nova Zelândia
  - Bayesian Knowledge Discovery
  - Algoritmos diversos, tais como C4.5, CN2 entre outros
- Linguagens com suporte a MD

# MineSet

- Ferramenta da Silicon Graphics para auxiliar processo de Mineração de Dados
- Possibilita visualização de dados multidimensionais
- Oferece utilização de algoritmos de mineração de dados e visualização gráfica dos modelos extraídos

Selection:

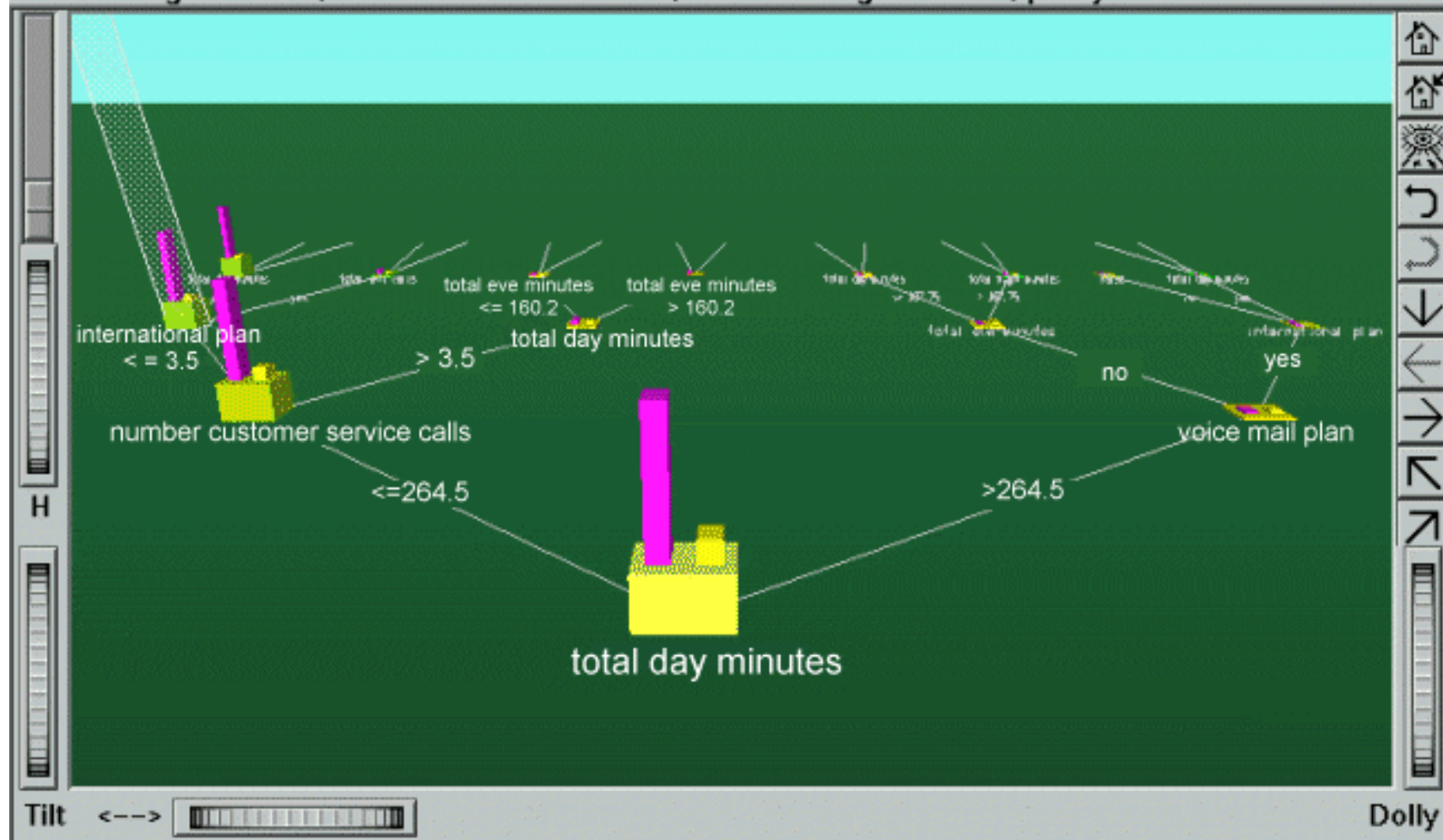
:total day minutes<= 264.45:number customer service calls<= 3.5:

Subtree weight:4309.00, test-set error:3.82+-0.51, test-set weight:1438.00, purity:



Pointer is over:

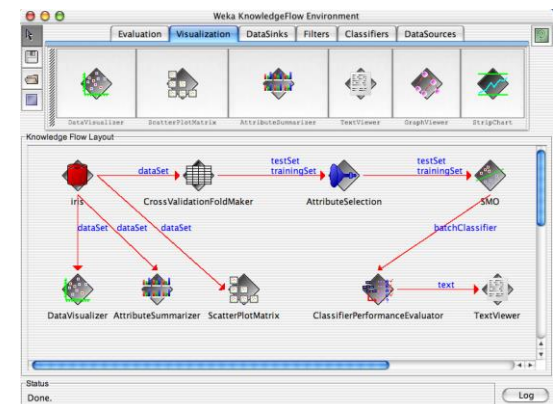
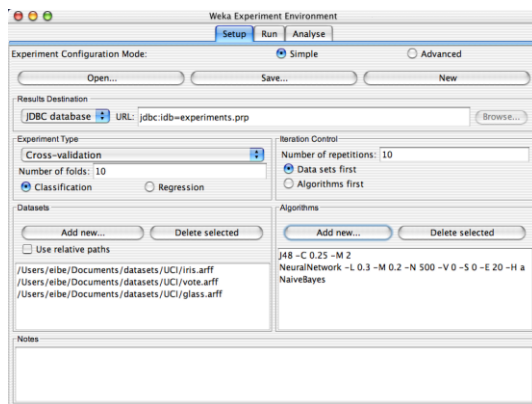
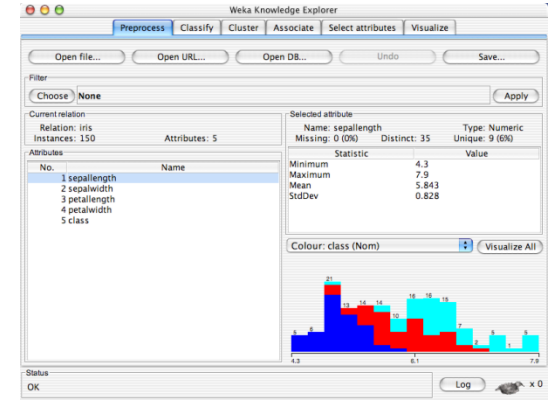
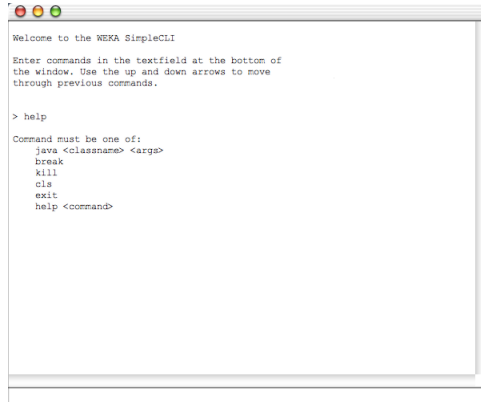
Subtree weight:5000.00, test-set error:5.46+-0.56, test-set weight:1667.00, purity:41.21



churned **False** **True**

Test-set error **low (0.00)** **medium (6.46)** **high (100.00)**

# WEKA





*“All things good to know  
are difficult to learn”*

*~ Greek Proverb ~*

- Material baseado em:
  - Notas Didáticas: Profa. Huei Diana Lee
  - Notas Didáticas: Profa. Maria Carolina Monard e Ronaldo Cristiano Prati.
  - Notas Didáticas: Prof. Walter Nagai
  - Notas Didáticas: Prof. E. Keogh
  - Notas Didáticas: Prof. Nitin Patel
  - Material IBM Research Brazil: Prof. Claudio Pinhanez