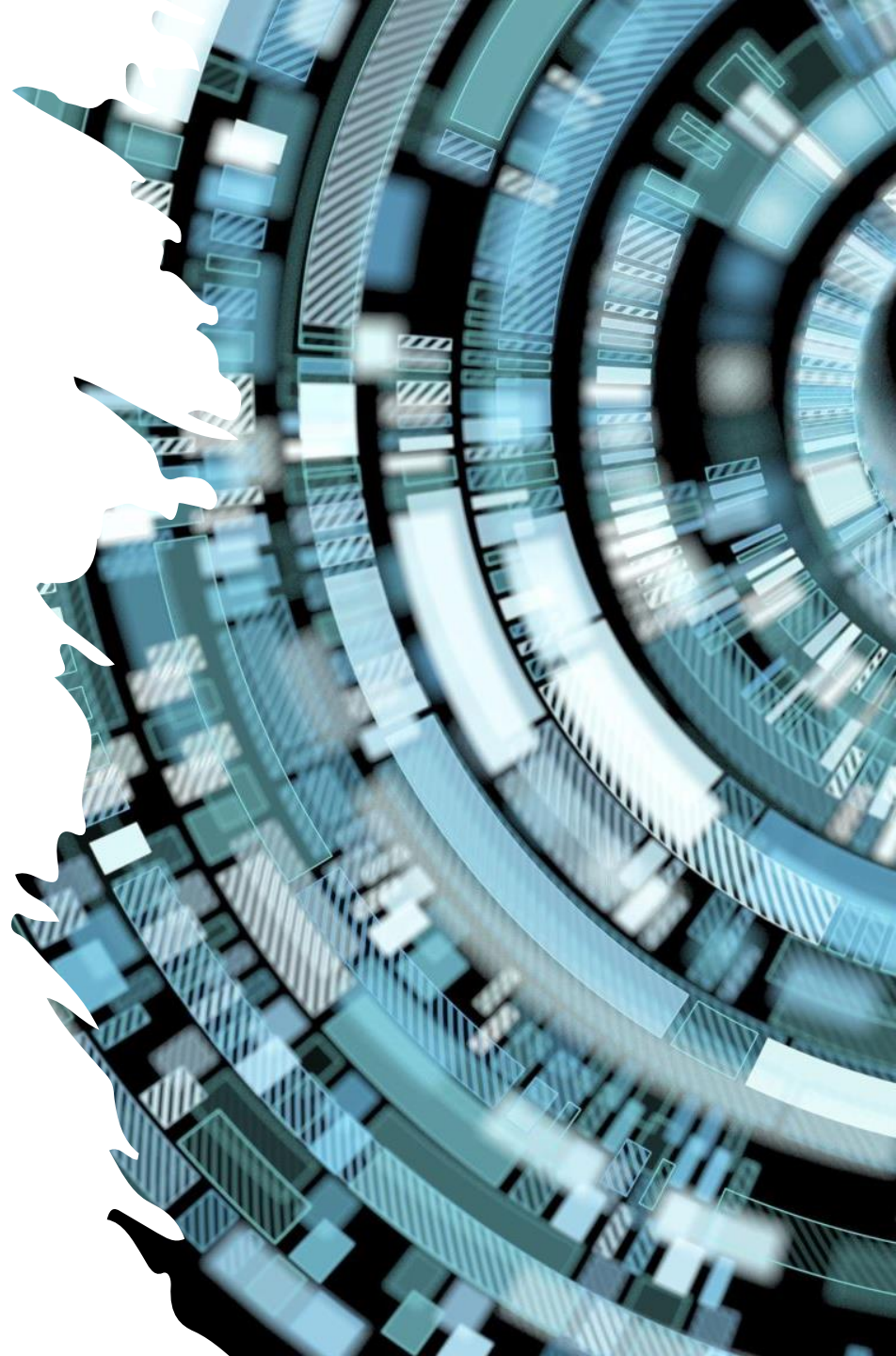


Extração de Conhecimento de Bases de Dados

*Knowledge Discovery in
Databases (KDD)*

Huei Diana Lee

Inteligência Artificial
CECE/UNIOESTE-FOZ



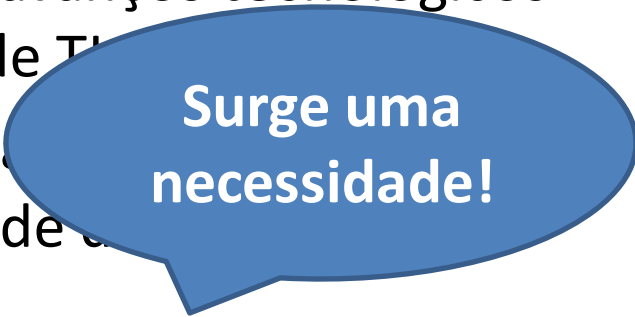
Motivação

Passado

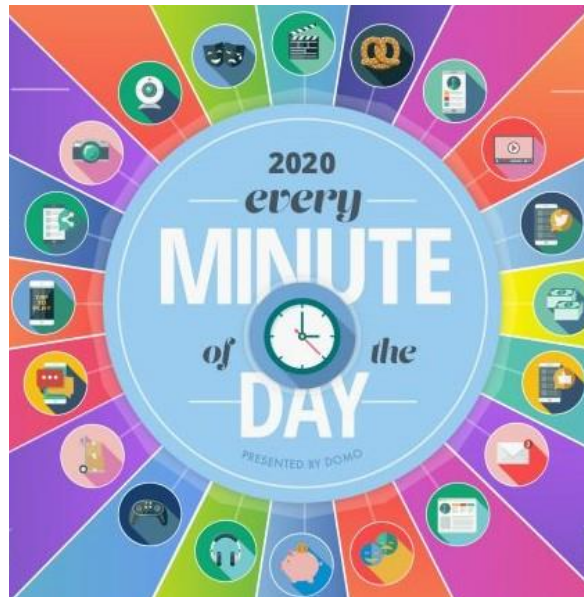
- Tecnologia limitada
- Armazenamento de pequenos volumes de dados (Mbytes)
- Consultas aos Dados
- Não existiam ferramentas para auxiliar a análise das informações obtidas

Presente/Futuro

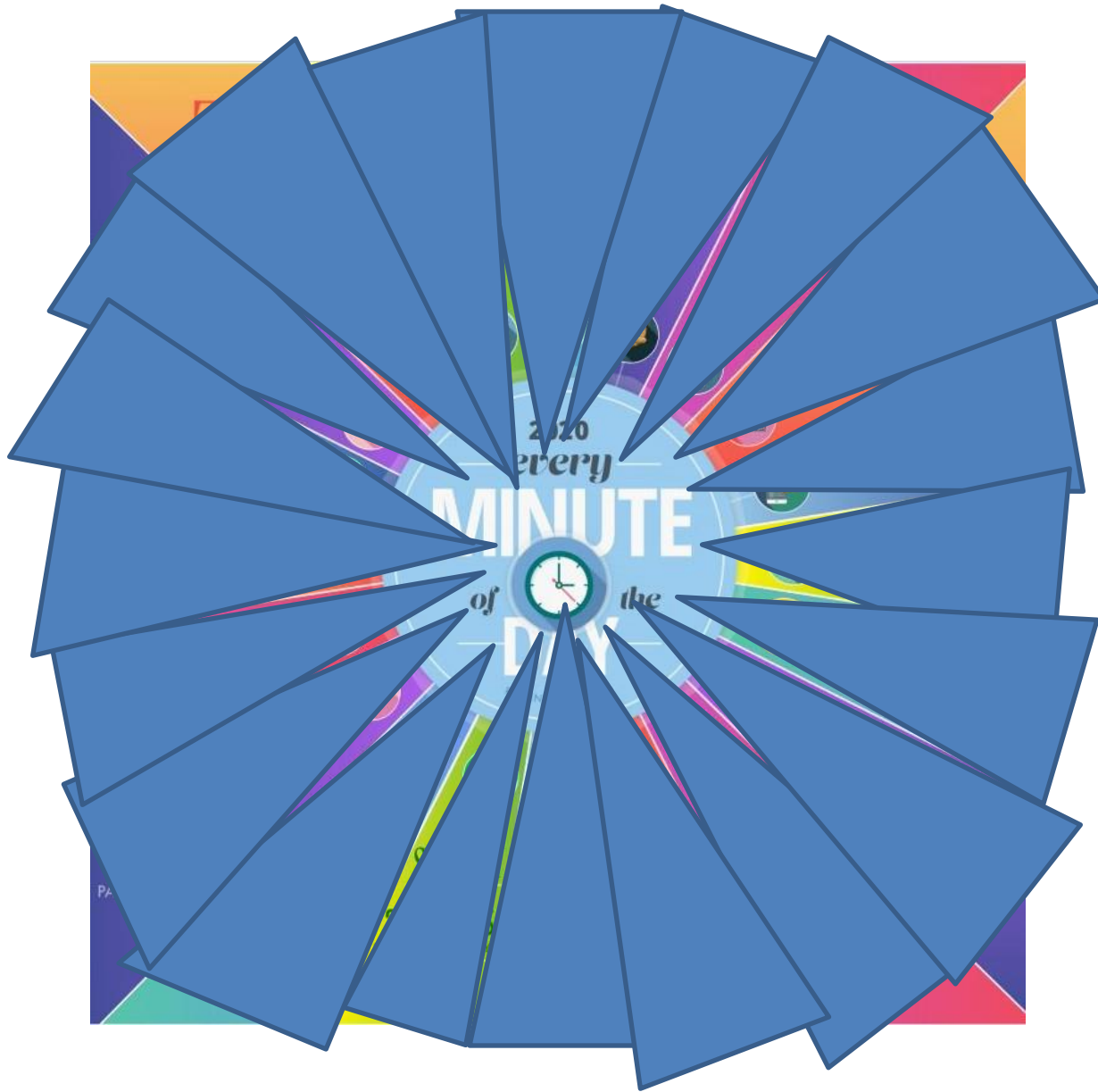
- Grandes avanços tecnológicos na área de TI
- Armazenamento de grandes volumes de dados (Tbyte... Pbyte...)
- Necessidade de conhecer e entender a BD
- O conhecimento extraído de uma BD deve ser usado para auxiliar as tomadas de decisões

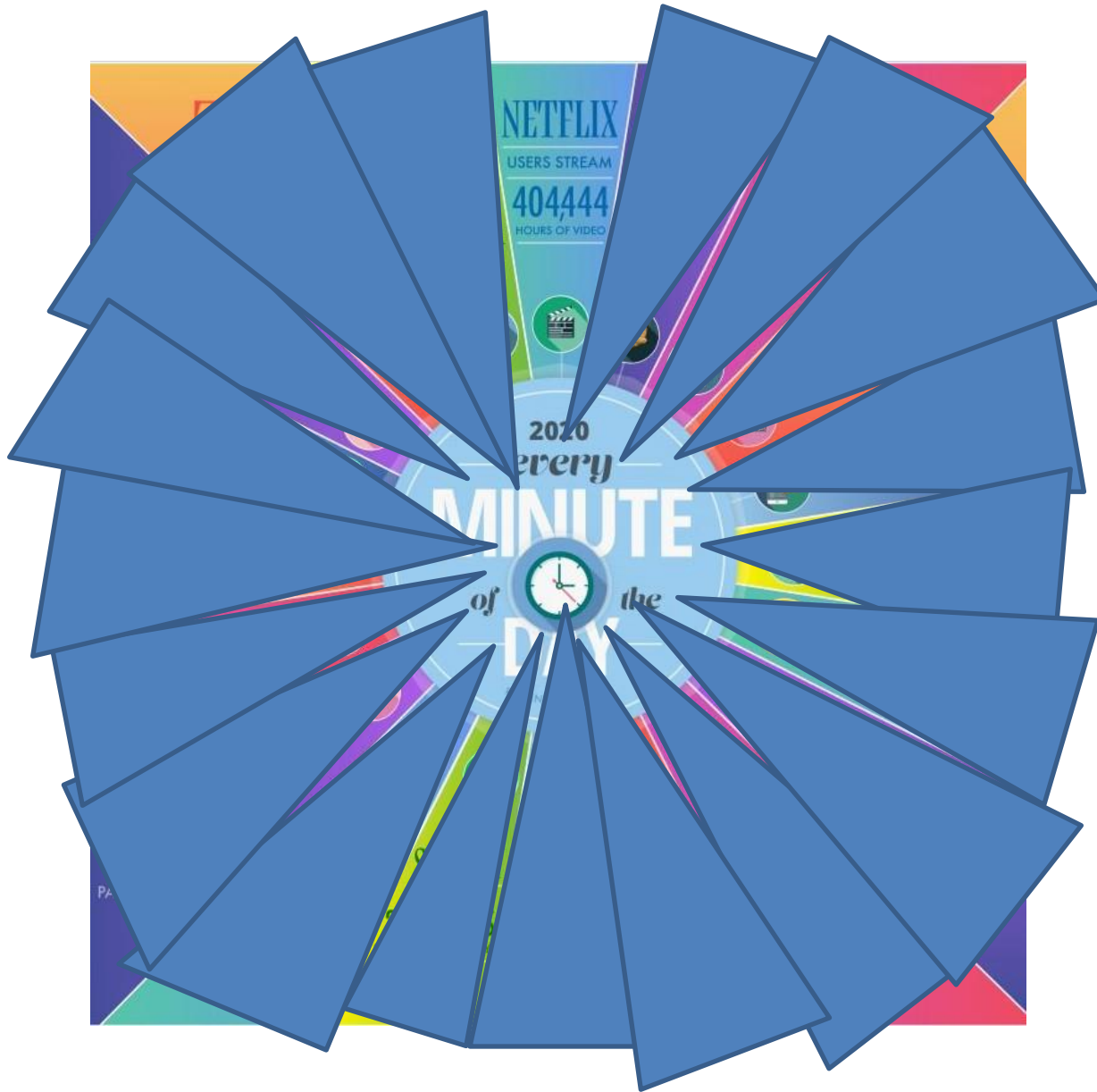


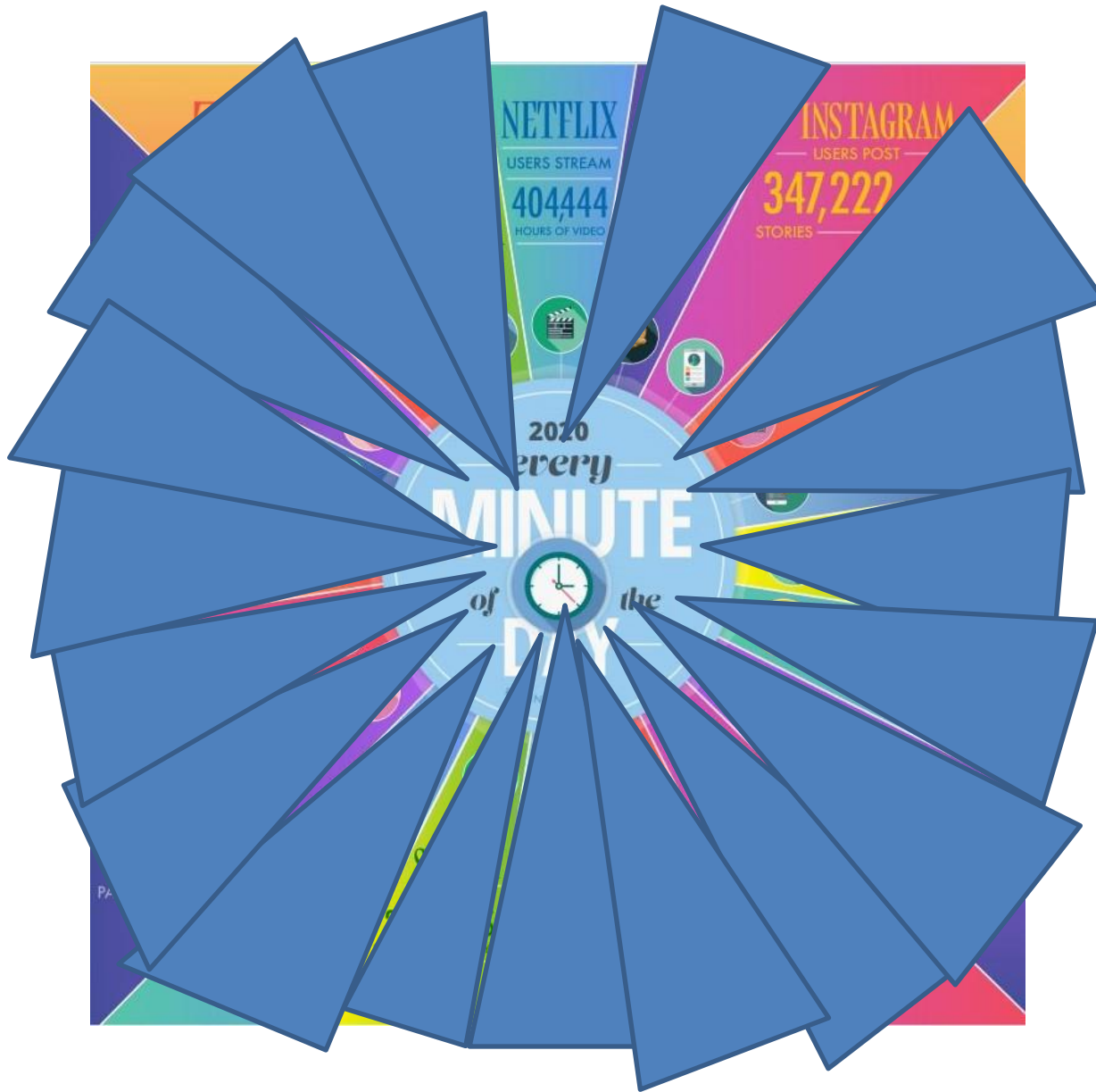
**Surge uma
necessidade!**

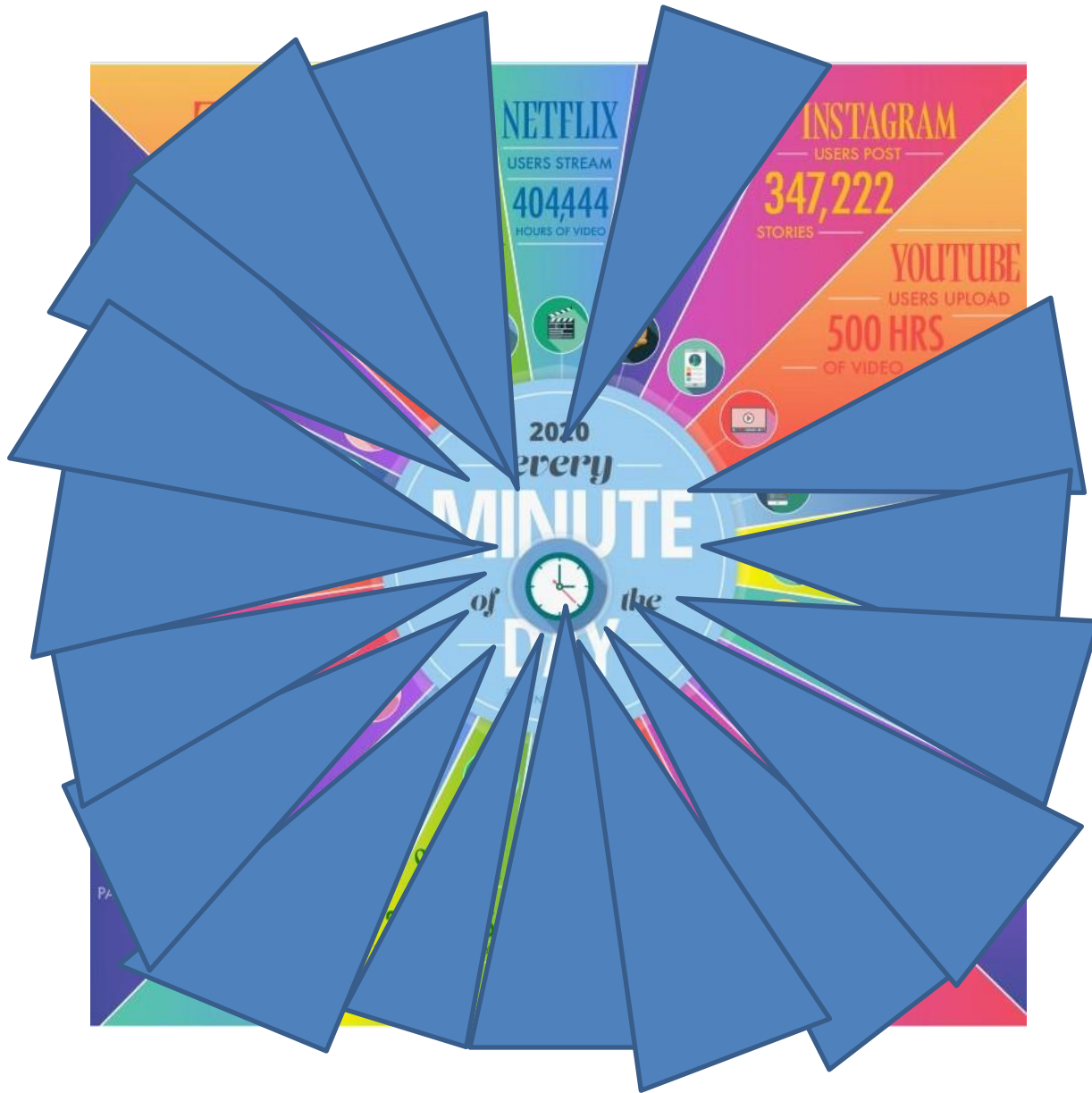


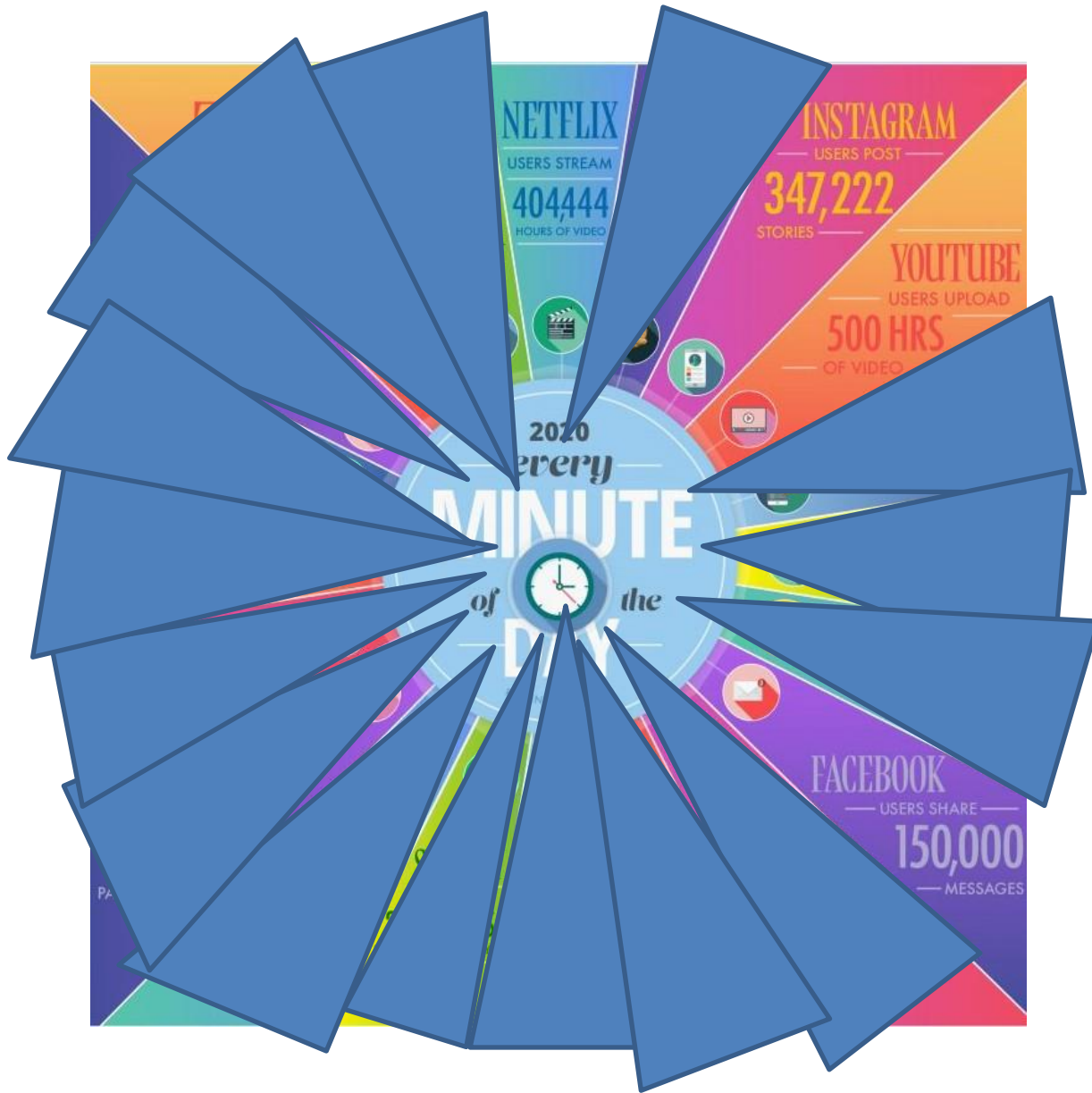
<https://www.domo.com/learn/data-never-sleeps-8>

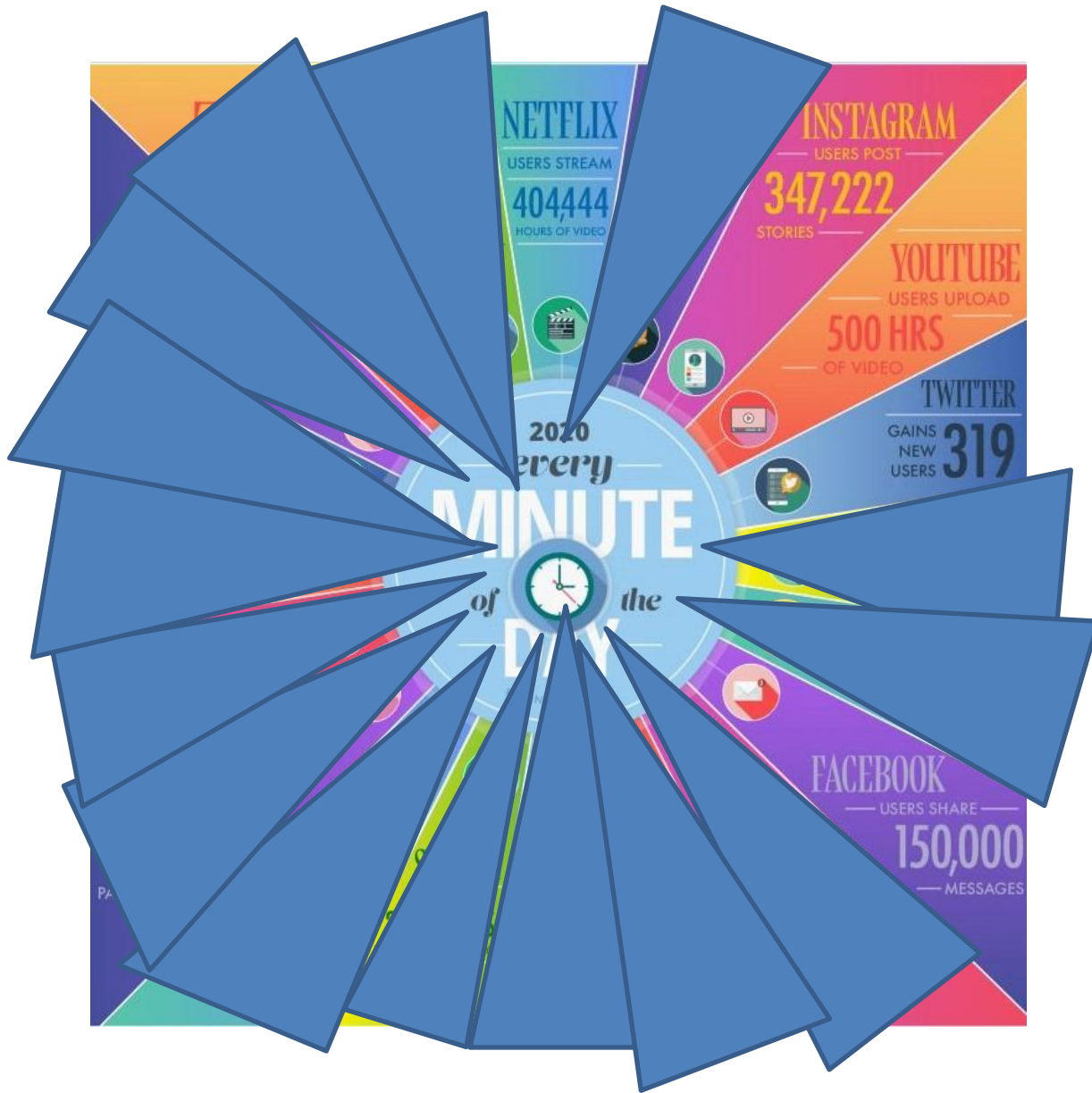


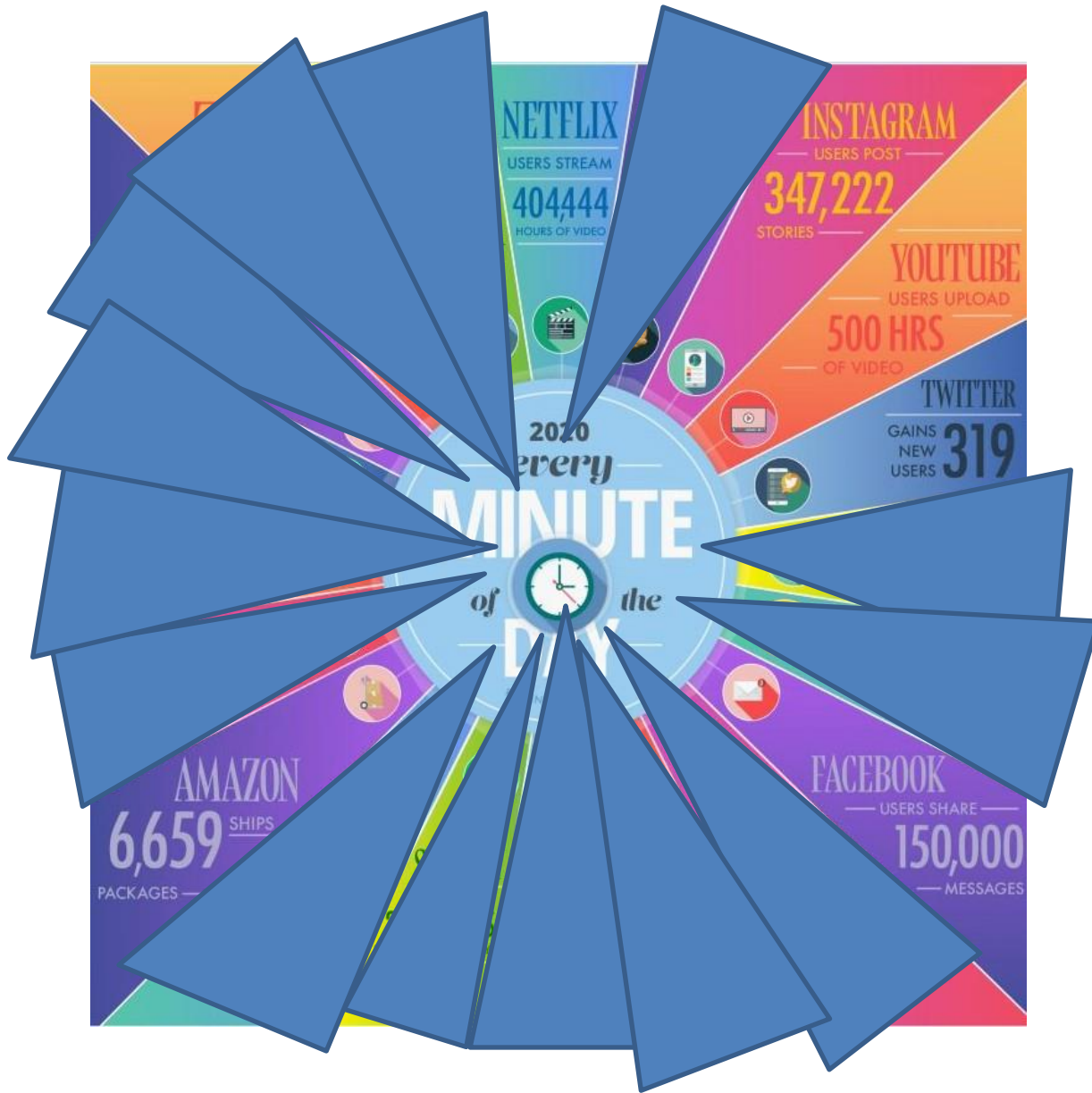


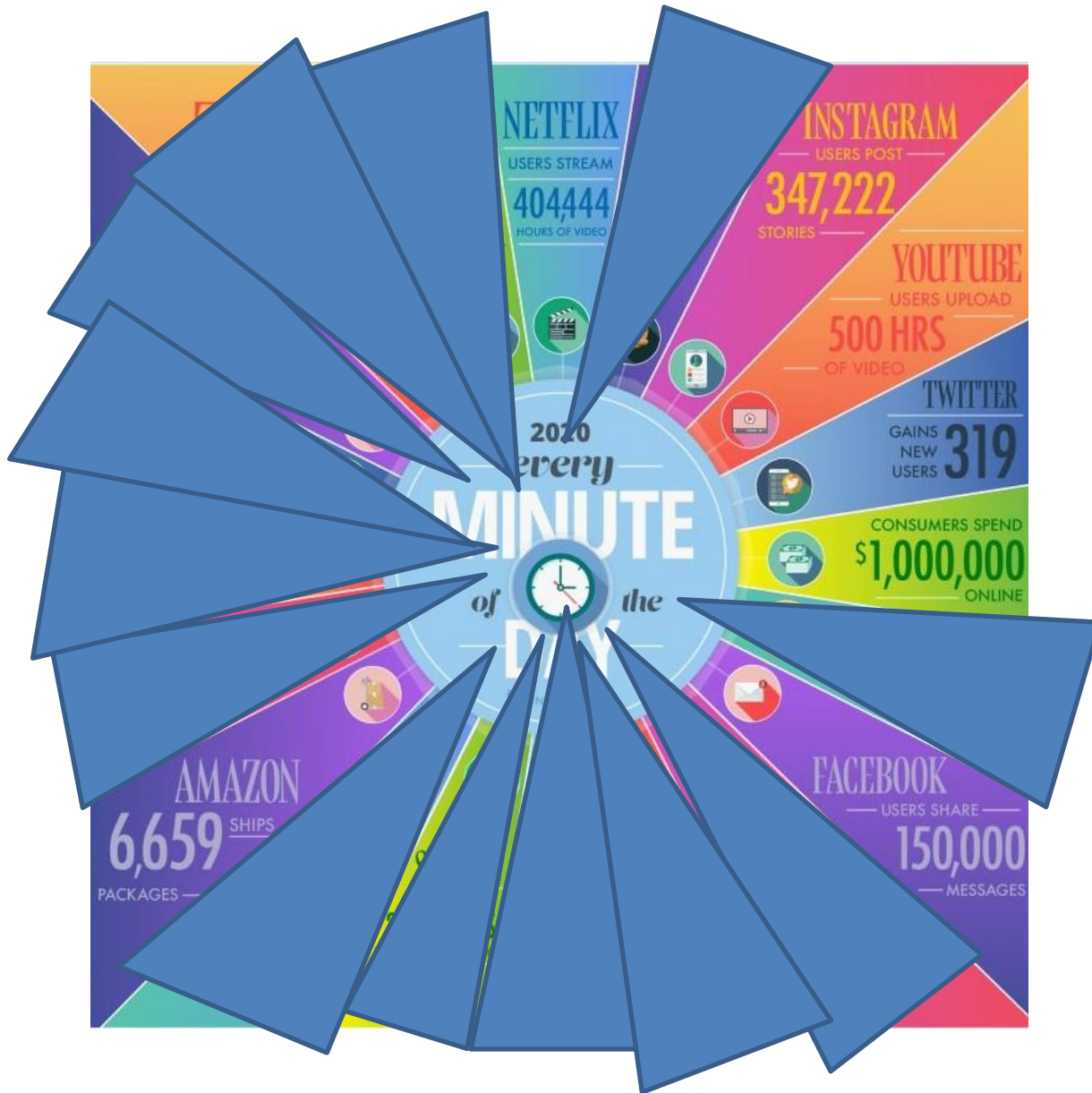


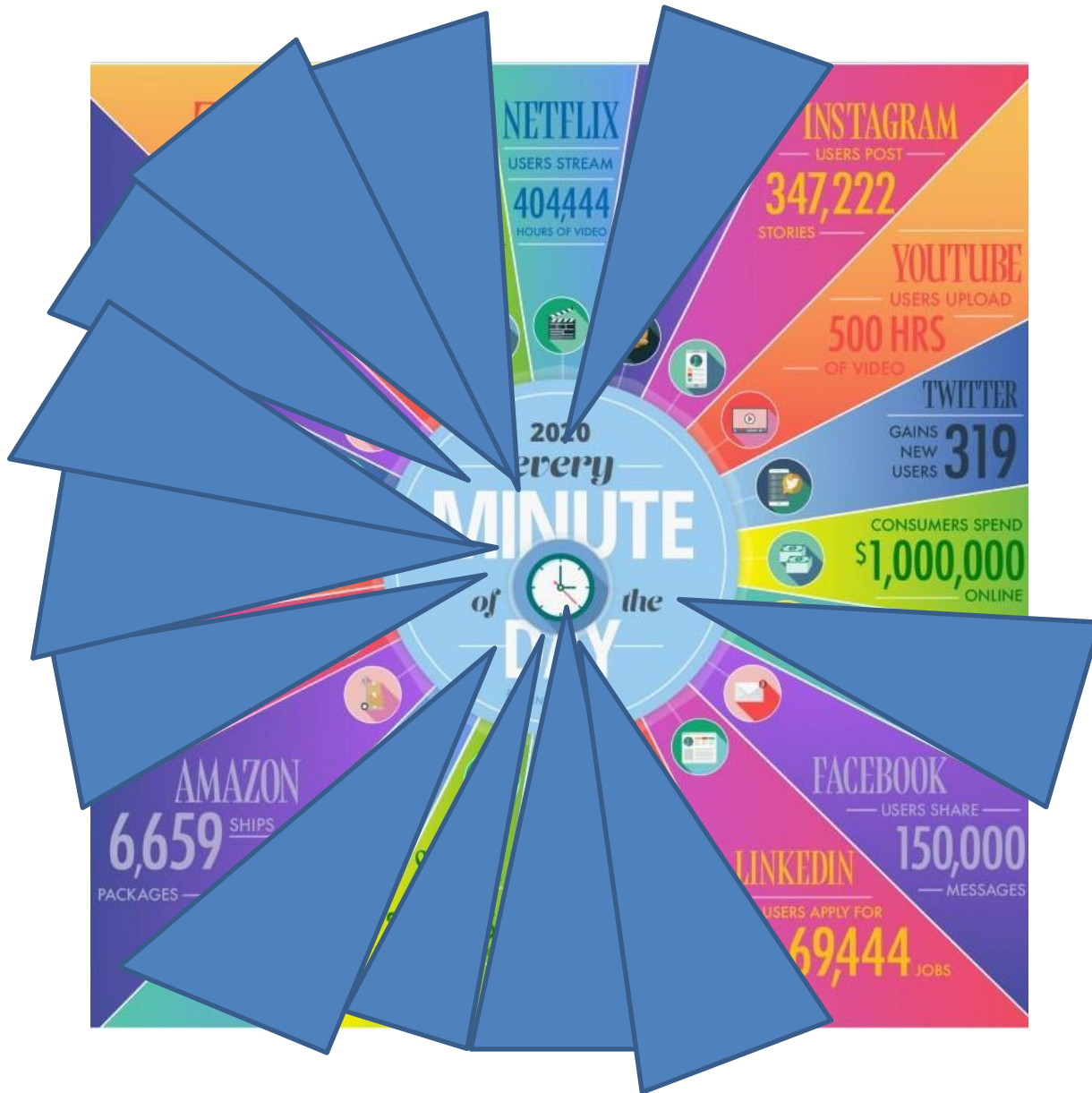


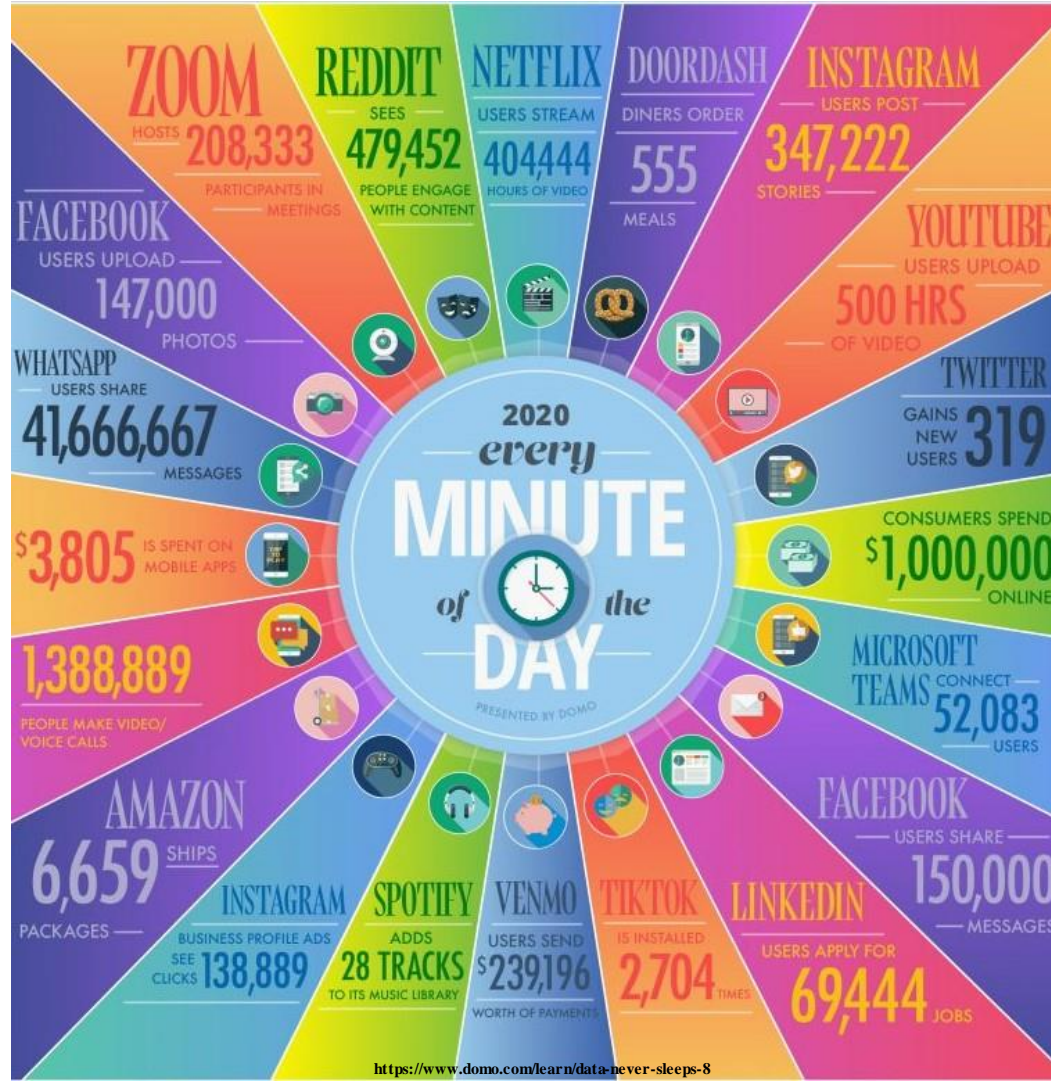












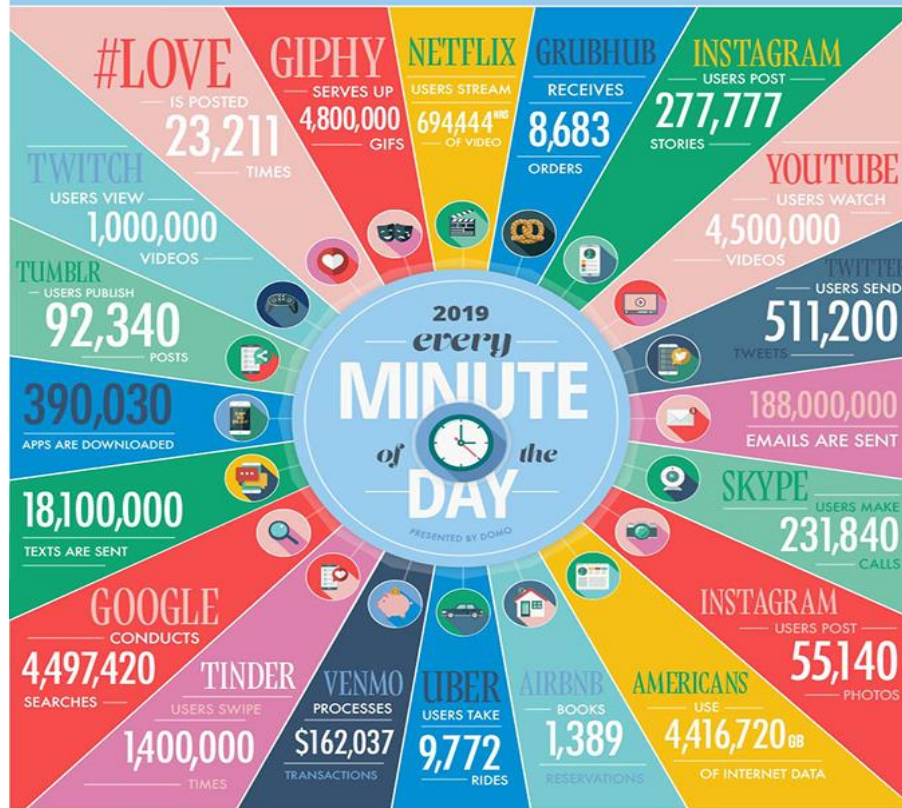
<https://www.domo.com/learn/data-never-sleeps-8>



DATA NEVER SLEEPS 7.0

How much data is generated *every minute*?

There's no way around it: big data just keeps getting bigger. The numbers are staggering, and they're not slowing down. By 2020, there will be 40x more bytes of data than there are stars in the observable universe. In our 7th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute—and the numbers are staggering.



The world's internet population is growing significantly year-over-year. As of January 2019, the internet reaches 56.1% of the world's population and now represents 4.39 billion people—a 9% increase from January 2018.



GLOBAL INTERNET POPULATION GROWTH 2012-2018 (IN BILLIONS)

SOURCES: STATISTA, INTERNET LIVE STATS, EXPANDED RAMBLINGS, NATIONAL ASSOCIATION OF CITY TRANSPORTATION OFFICIALS, WIRE

The ability to make data-driven decisions is crucial to any business. With each click, swipe, share, and like, a world of valuable information is created. Domo puts the power to make those decisions right into the palm of your hand by connecting your data and your people at any moment, on any device, so they can make the kind of decisions that make an impact.

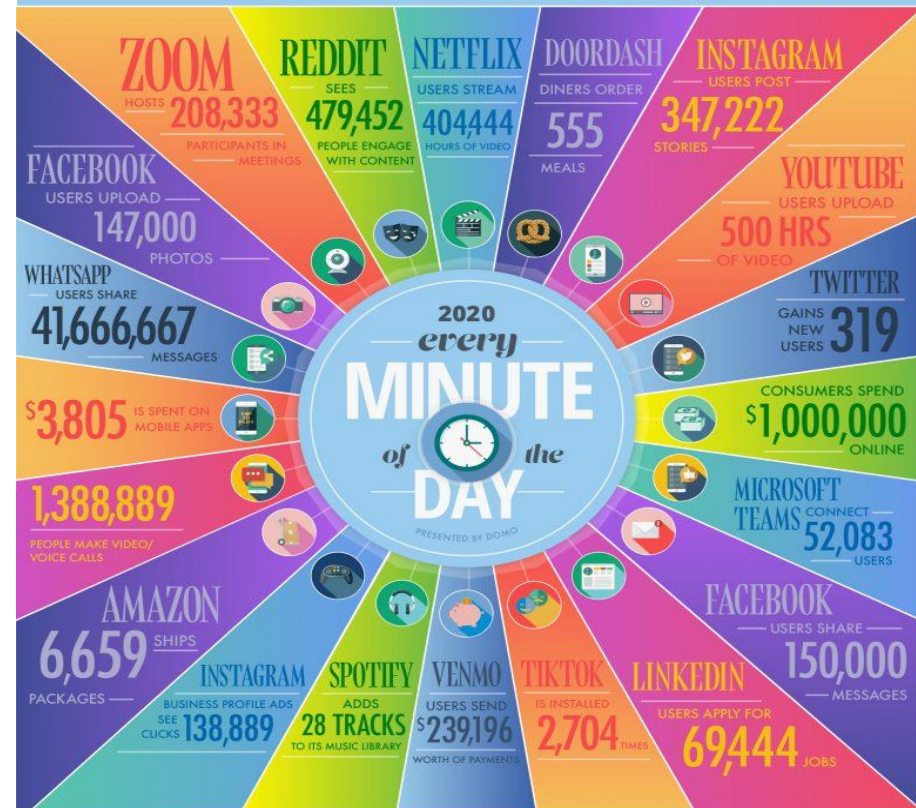
Learn more at domo.com



DATA NEVER SLEEPS 8.0

How much data is generated *every minute*?

In 2020, the world changed fundamentally—and so did the data that makes the world go round. As COVID-19 swept the globe, nearly every aspect of life—from work to working out—moved online, and people depended more and more on apps and the Internet to socialize, educate and entertain ourselves. Before quarantine, just 15% of Americans worked from home. Now over half do. And that's not the only big shift. In our 8th edition of Data Never Sleeps, we bring you the latest stats on how much data is being created in every digital minute—a trend that shows no sign of stopping.



The world's internet population is growing significantly year-over-year. As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people—a 6% increase from January 2019.



GLOBAL INTERNET POPULATION GROWTH 2014-2020 (IN BILLIONS)

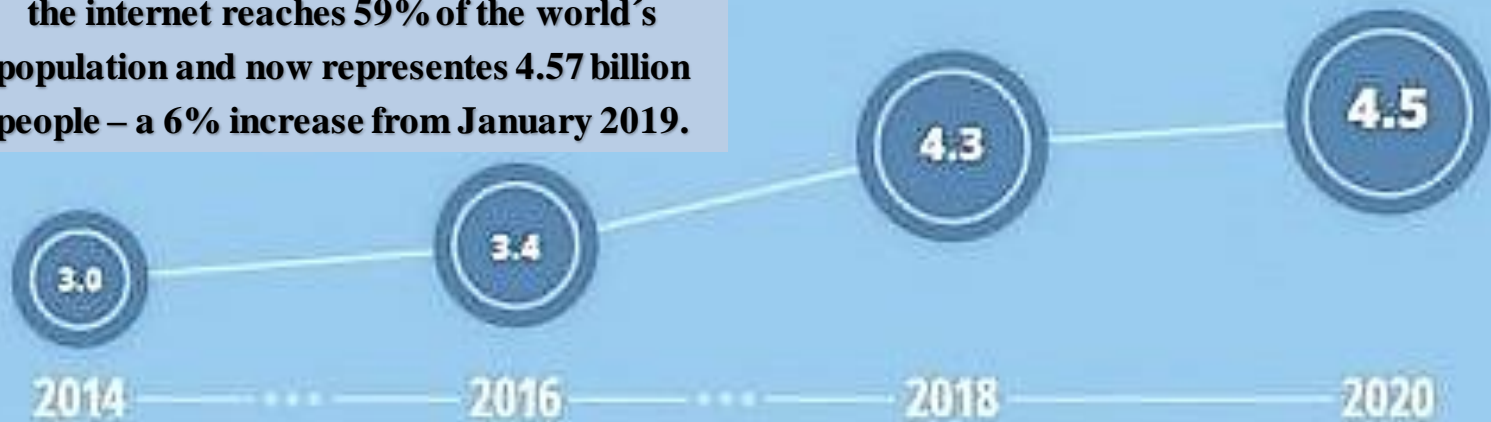
SOURCES: STATISTA, VISUAL CAPITALIST, BUSINESS INSIDER, GAMSPOT, TECHCRUNCH, OMNICORE AGENCY, DOORDASH, BUSINESS OF APPS, NEW YORK TIMES, MUSIC BUSINESS WORLDWIDE, INC., THE VERGE, INC., POKETWORK, DUSTIN STAFF, REDDIT, UBER, AMAZON, WIRE

As the world changes, businesses need to change with the times—and that requires data. Every click, swipe, share or like tells you something about your customers and what they want, and Domo is here to help your business make sense of all of it. Domo gives you the power to make data-driven decisions at any moment, on any device, so you can make smart choices in a rapidly changing world.

Learn more at domo.com



The world's internet population is growing significantly year over year. As of April 2020, the internet reaches 59% of the world's population and now represents 4.57 billion people – a 6% increase from January 2019.



GLOBAL INTERNET POPULATION GROWTH 2014-2020
(IN BILLIONS)

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

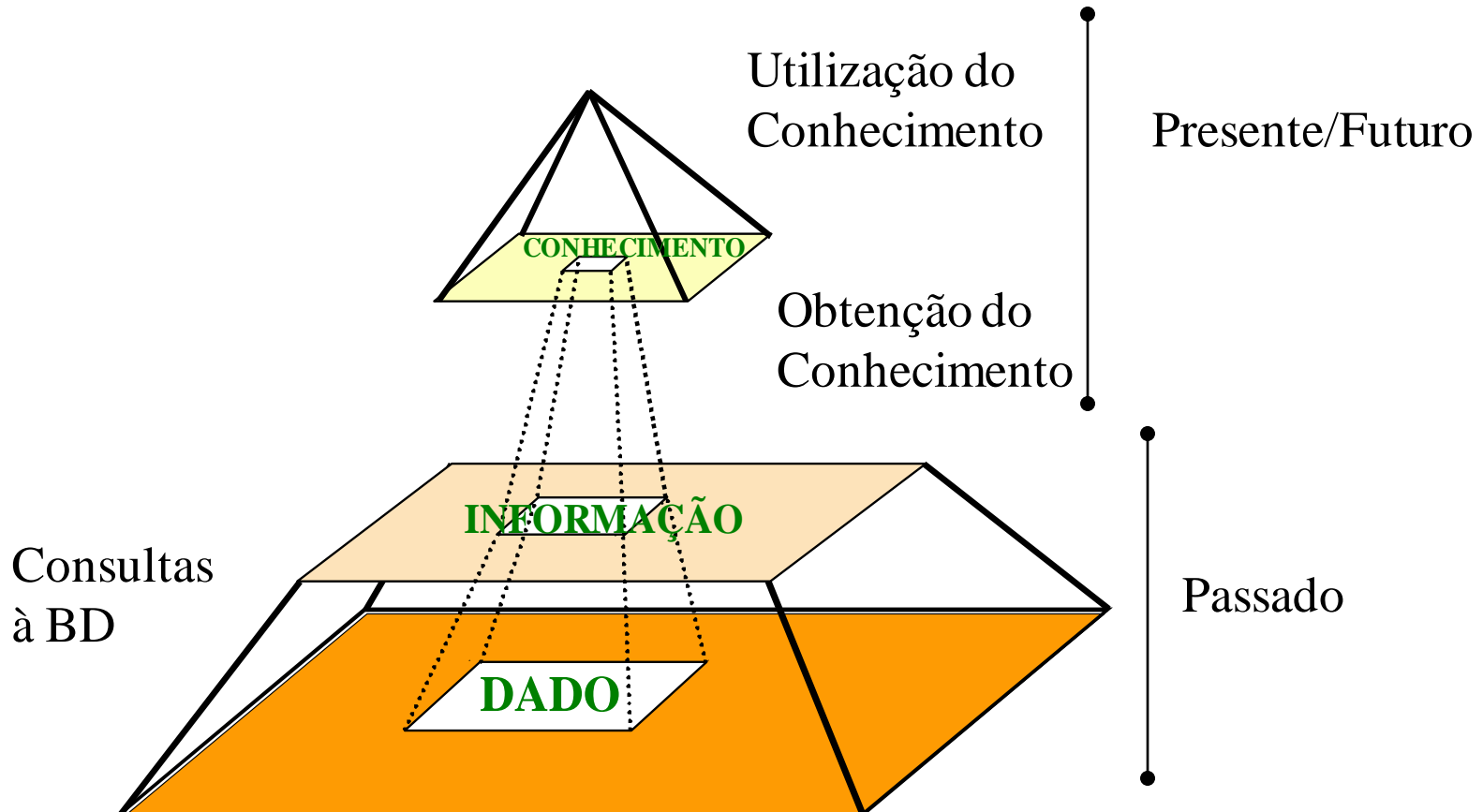
Gigantes, Monstros & “Leis”

O que é Big Data?

- **V**olume
- **V**ariedade
- **V**elocidade
- **V**eracidade

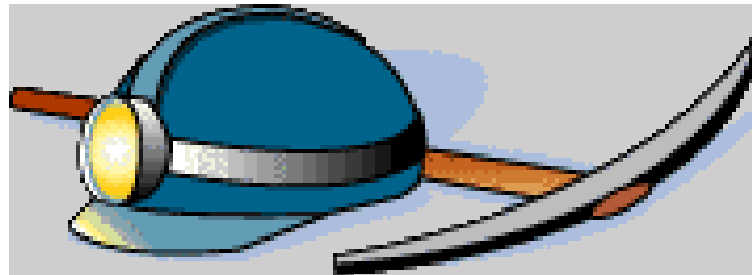
Motivação

Pirâmide do Conhecimento



Introdução

O objetivo da extração de conhecimento é descobrir situações anômalas e/ou interessantes, tendências, padrões e seqüências 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

- MasterCard:
 - identificar perfis de clientes
 - monitorar reações a campanhas publicitárias
 - identificar novas tendências e possíveis novos produtos



Exemplos de aplicações

- Dell:
 - Problema: 50% dos clientes da Dell encomendam computadores pelo site na web. Porém, a taxa de retenção é de 0,5% (visitantes da página que se tornam clientes)
 - Abordagem para Solução: Pela sequência de clicks, agrupar clientes e desenvolver o website de modo a maximizar o número de clientes que eventualmente comprarão
 - Benefício: Aumento de vendas

Exemplos de aplicações

- Sistemas de recomendação:
 - Oportunidade de negócio: Usuários avaliam itens na web. Como usar essa informação de outros usuários para inferir avaliações para um usuário em particular?
 - Solução: Usar filtragem colaborativa
 - Benefício: Aumento na rentabilidade por meio de “cross selling”



DeepFAMA: High-Quality, High Volume Short Text Classifier

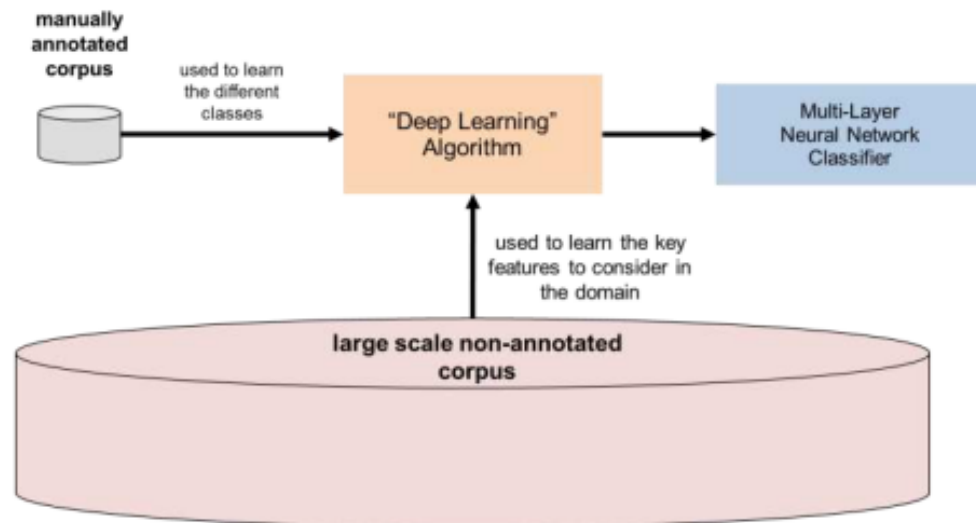
- *DeepFAMA* is a **short text classifier** developed by IBM Research – Brazil
- Applicable to conversational short texts (social media, SMS, call-center transcripts)
- Available in **English** and **Portuguese**;
- Implementation available for **high volume**, real time production scenarios (IBM Streams)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



"FAMA"

Greek goddess of gossip and rumor

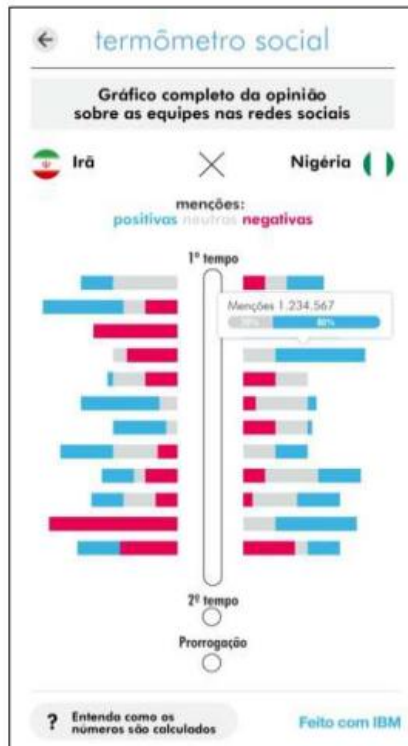
New Deep Convolutional Neural Network (DCNN) Algorithm



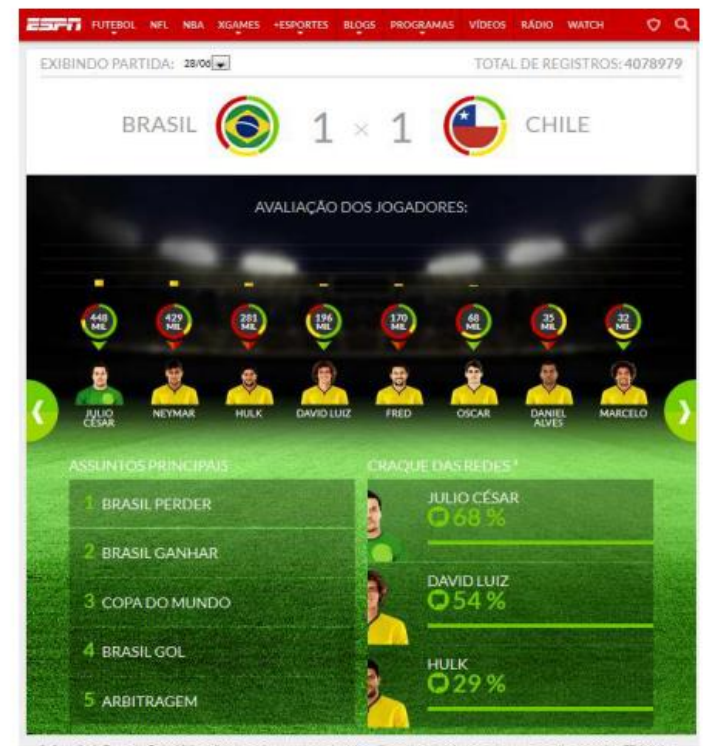
Cicero N. dos Santos and Maira Gatti. *Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts*. **Proceedings of COLING 2014**, pages 69–78, Dublin, Ireland, August 23-29 2014.

© 2015 IBM Corporation

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

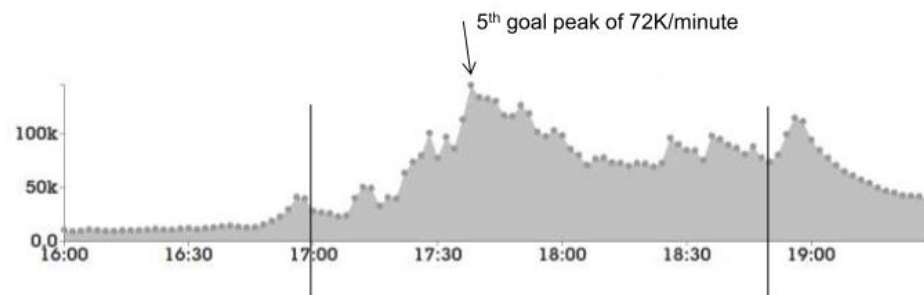


Globo 2nd screen app
1.4M downloads, 1.8M page views

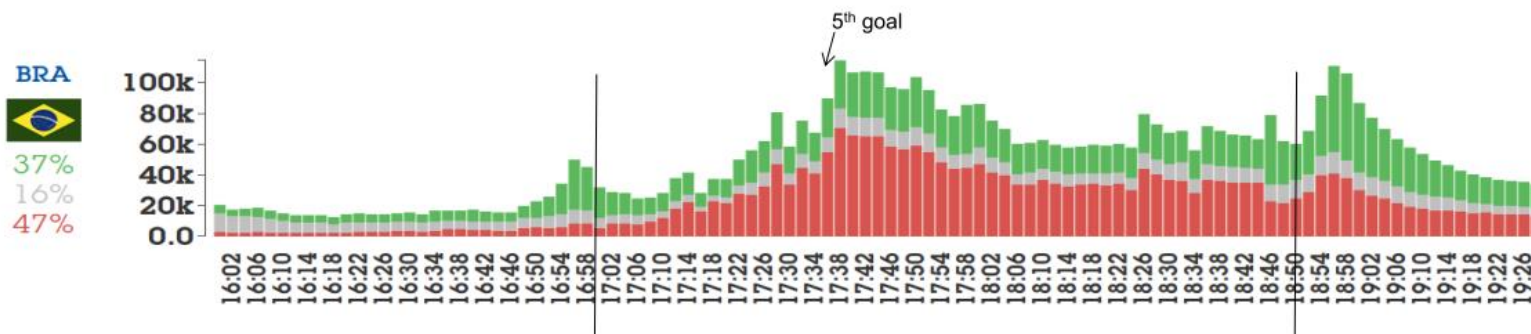


ESPN Brazil
54.3K page views

BRA 1x7 GER: Largest Event in SN History



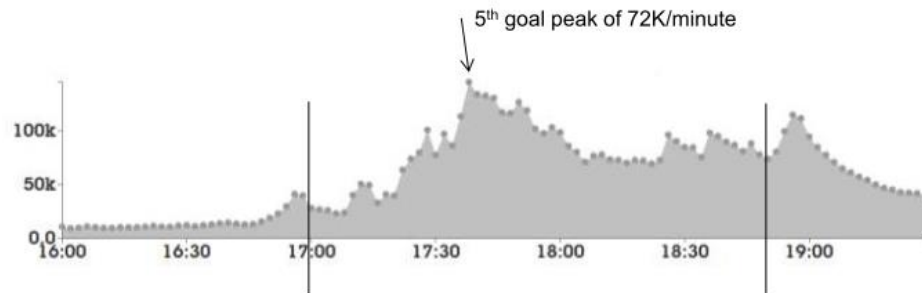
- globally 35.6M tweets (WR)
- 6.8M posts in Portuguese (19% of world)
- peak of 72K/minute
- 1.4M tweets after the game



52

©Copyright 2014 IBM Co.

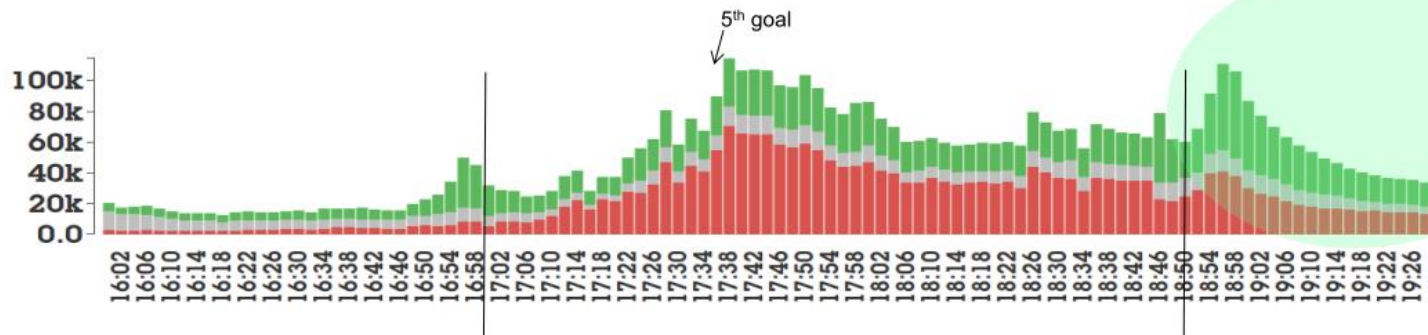
BRA 1x7 GER: Largest Event in SN History



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

BRA

 37%
 16%
 47%



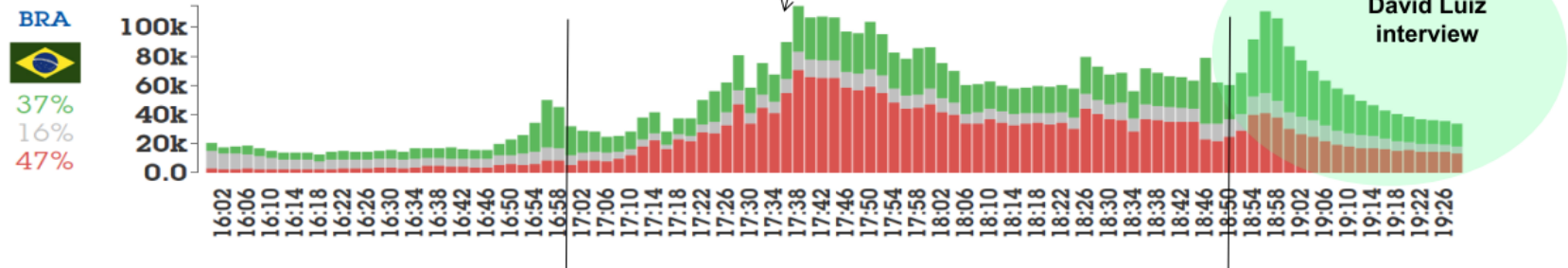
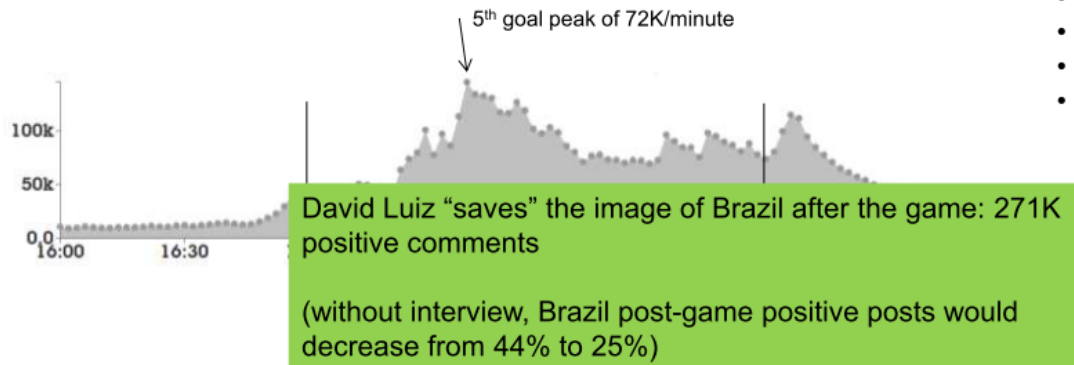
53

©Copyright 2014 IBM Co.

BRA 1x7 GER: Largest Event in SN History



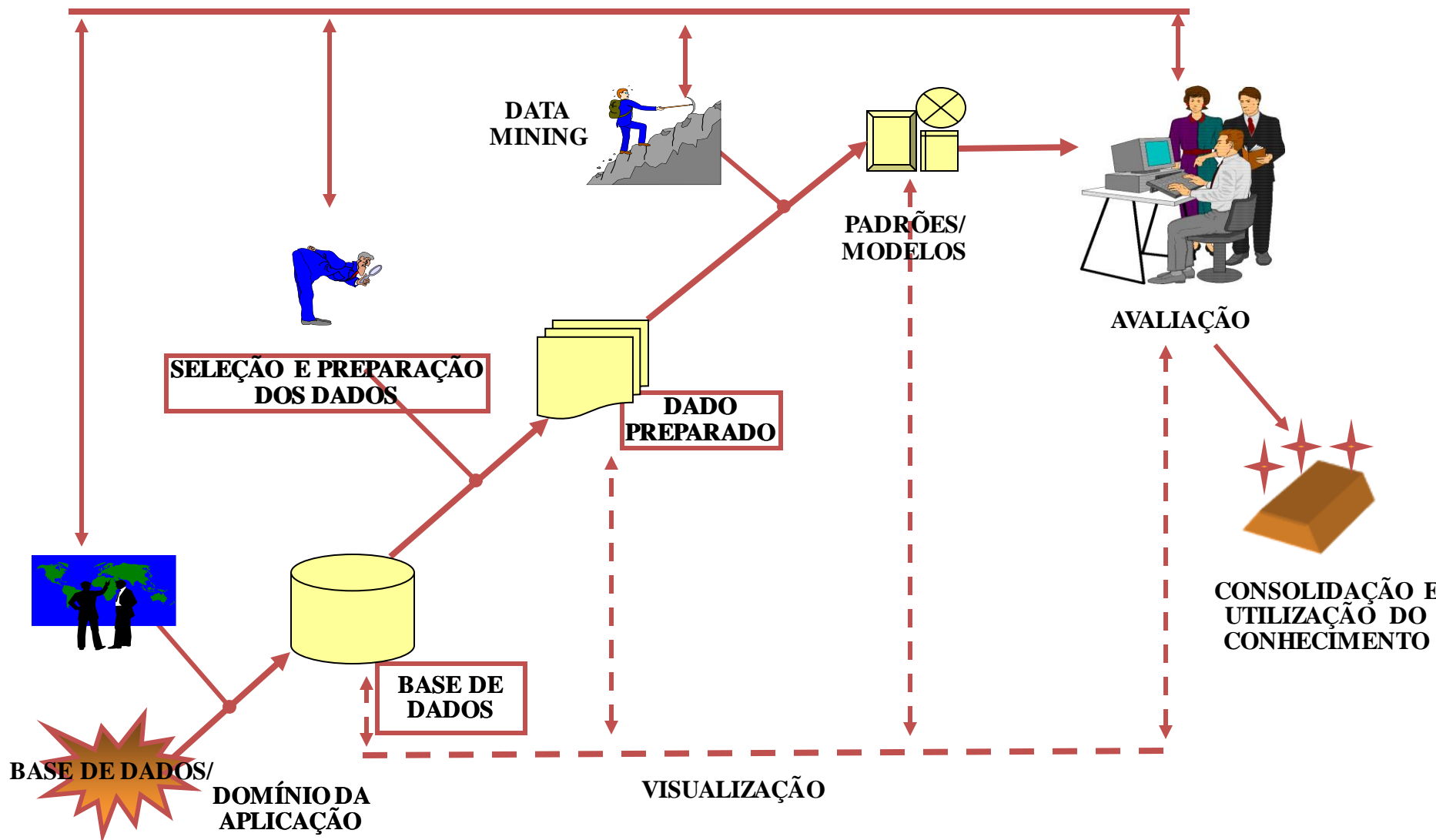
- globally 35.6M tweets (WR)
- 6.8M posts in Portuguese (19% of world)
- peak of 72K/minute
- 1.4M tweets after the game



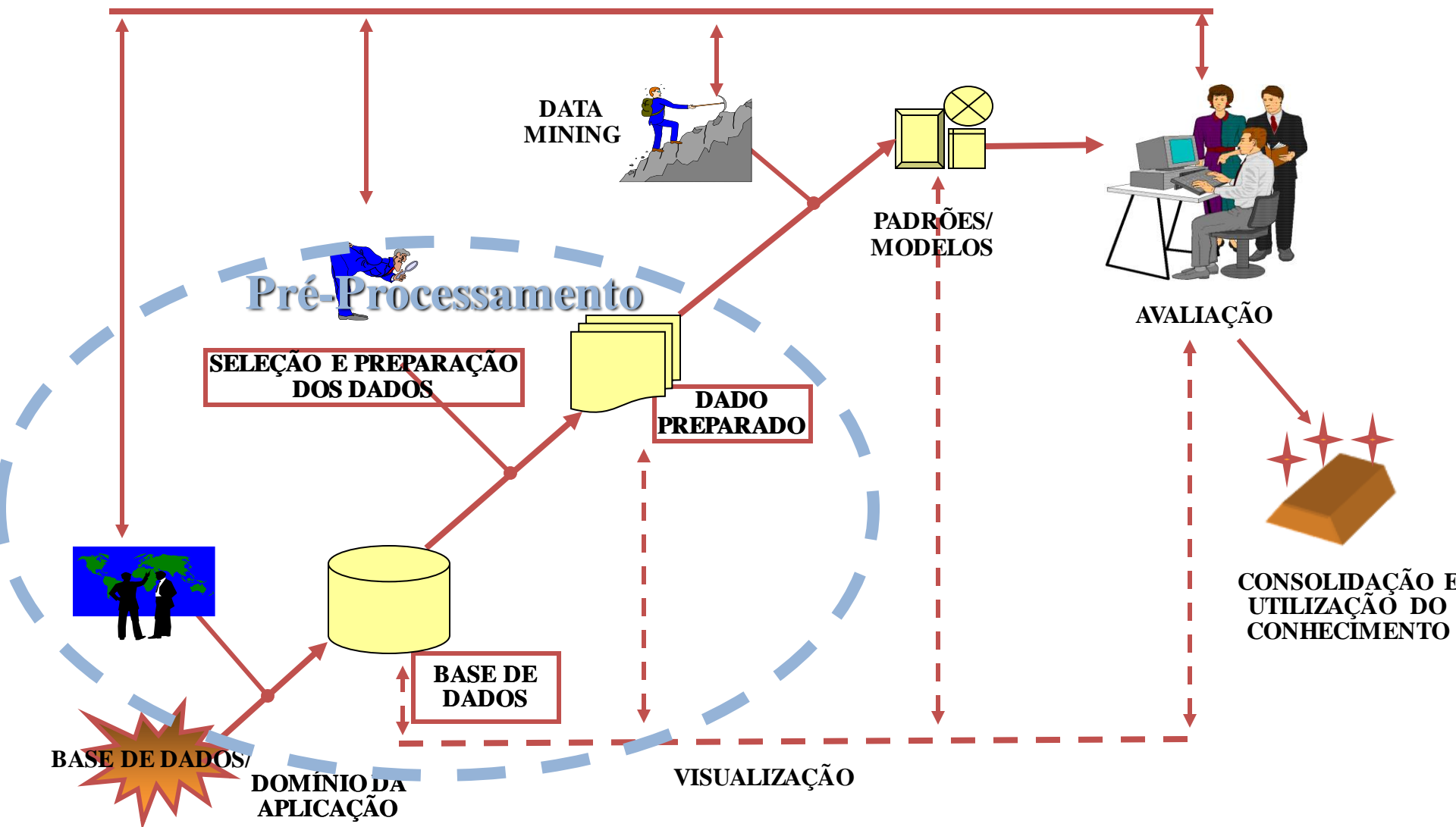
54

©Copyright 2014 IBM Co.

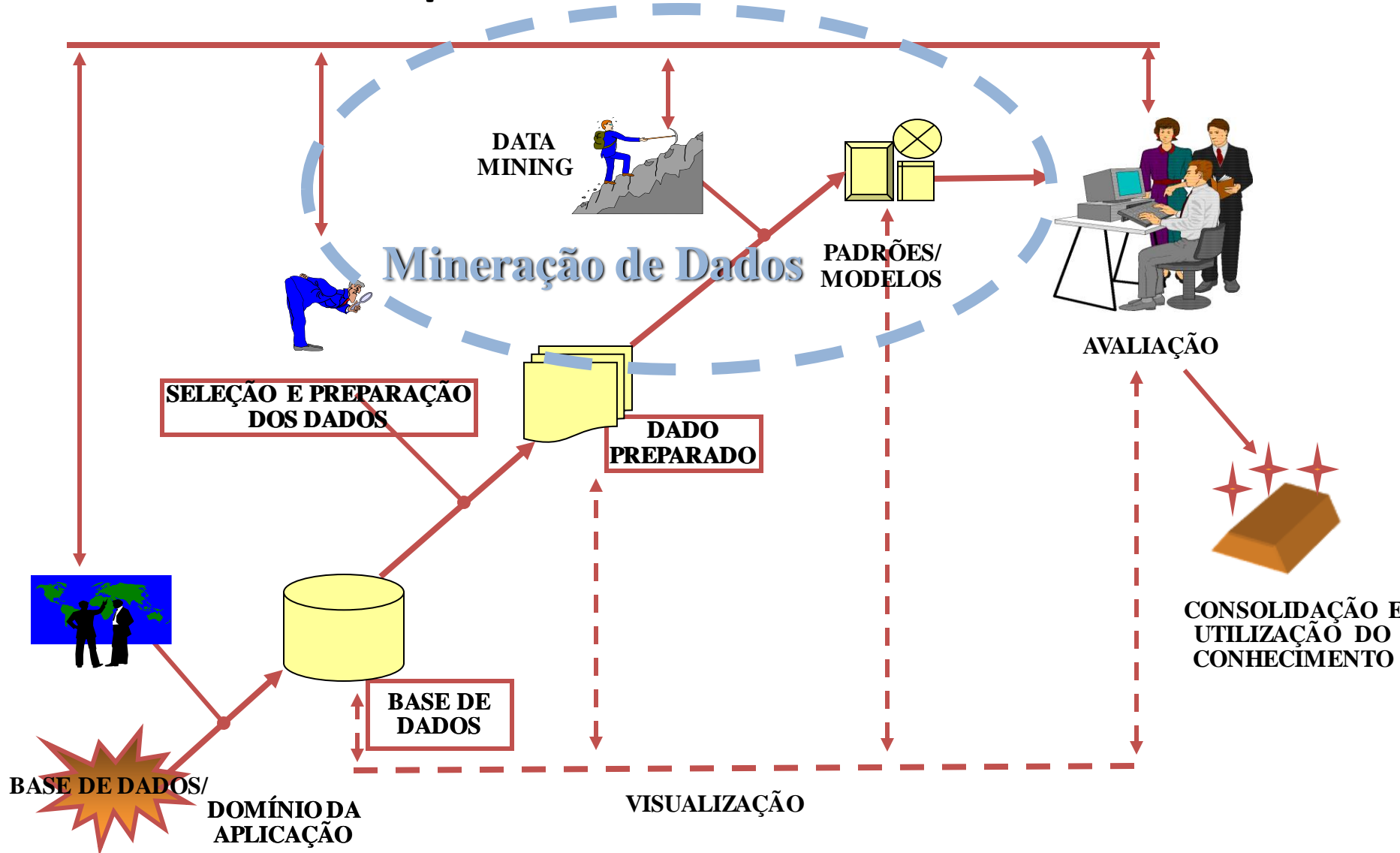
Etapas do Processo KDD



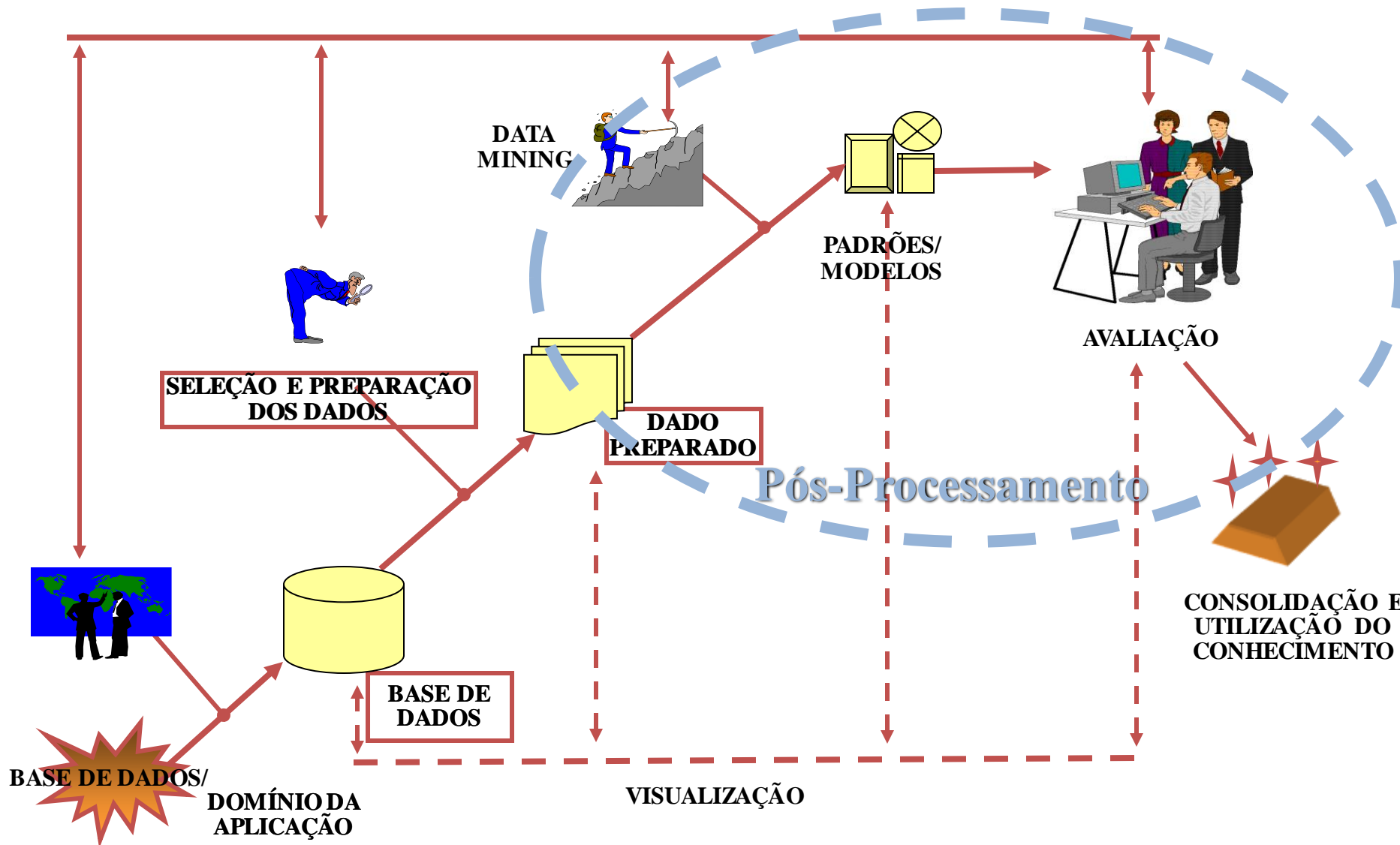
Etapas do Processo KDD



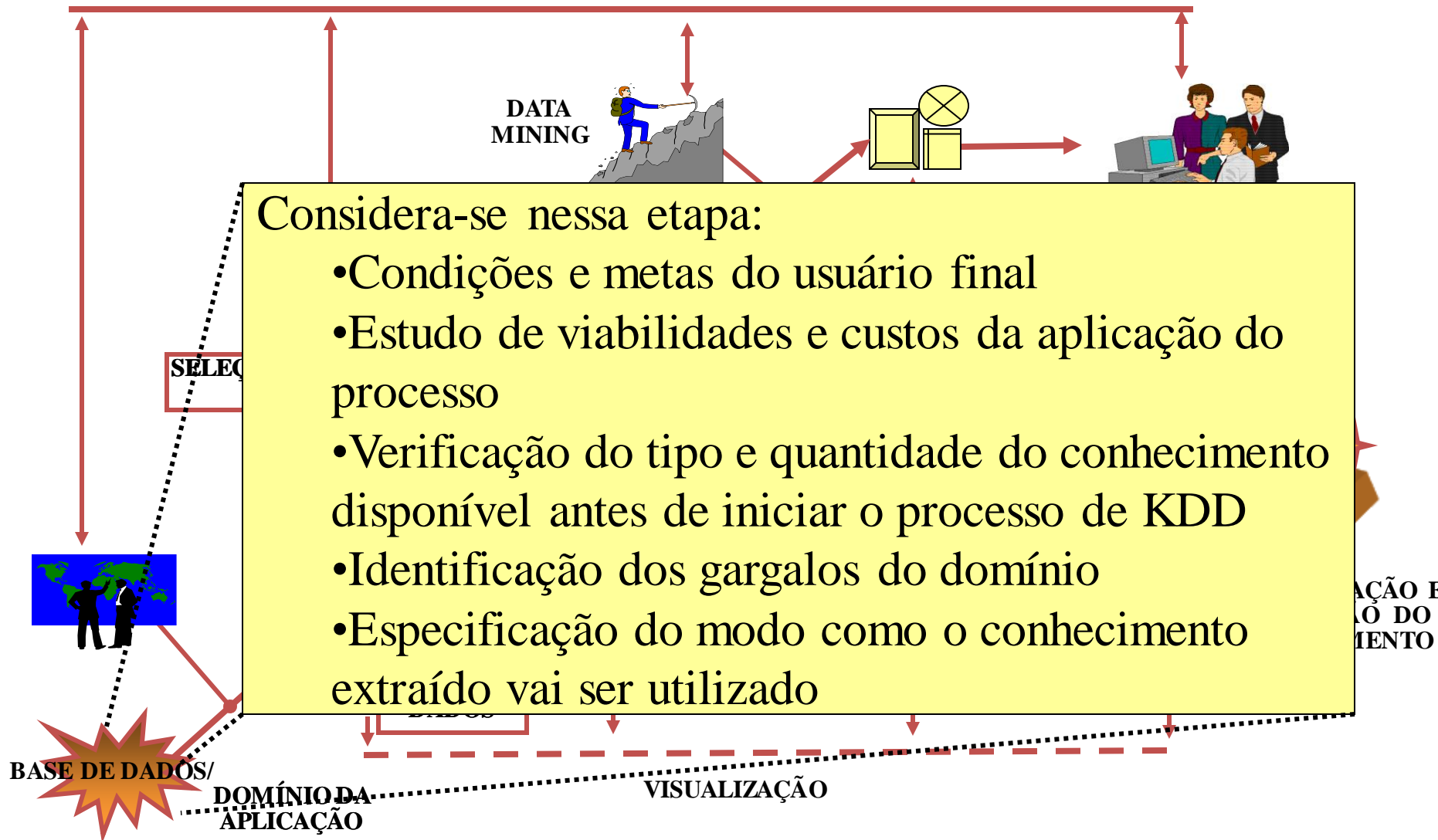
Etapas do Processo KDD



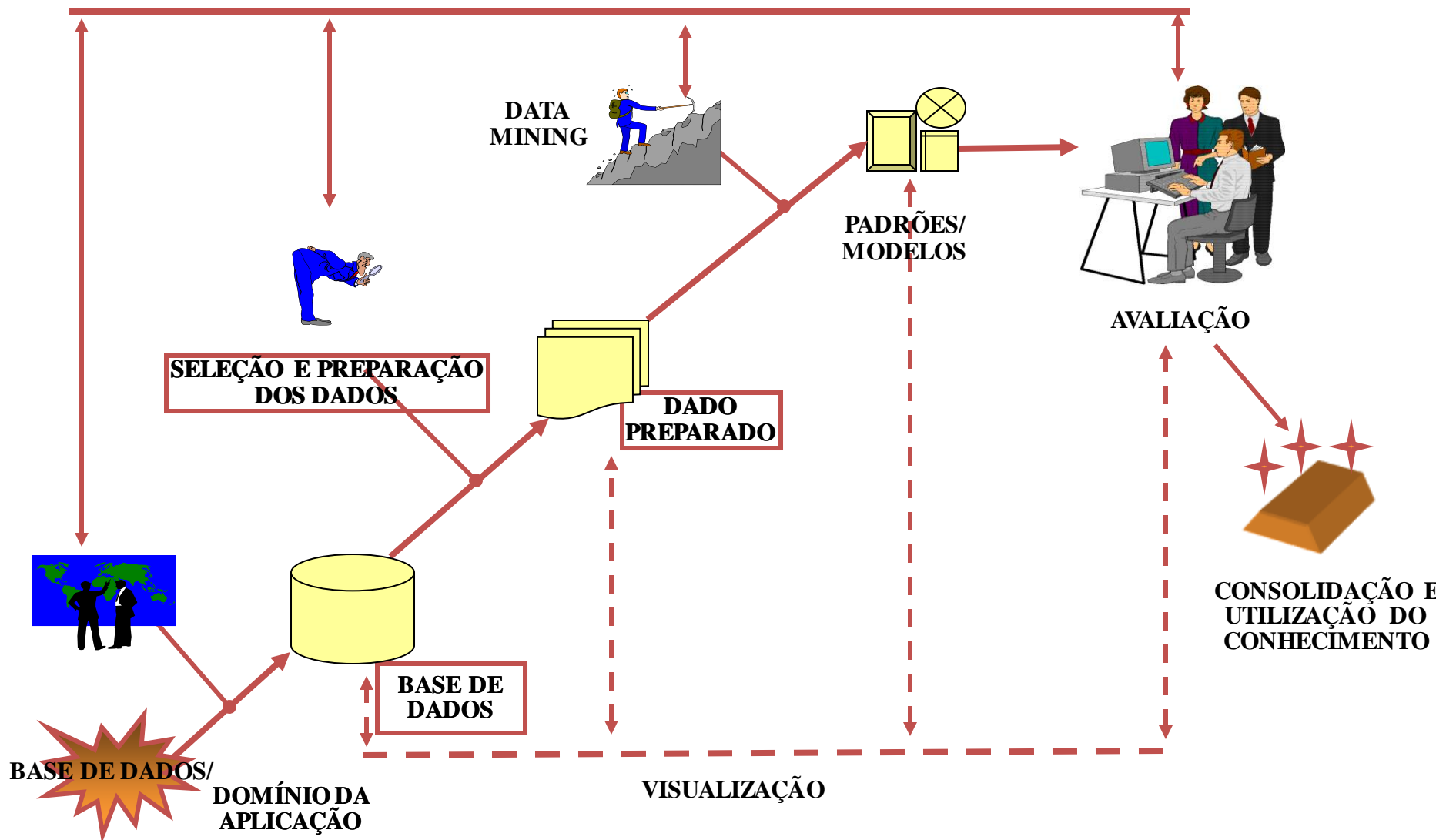
Etapas do Processo KDD



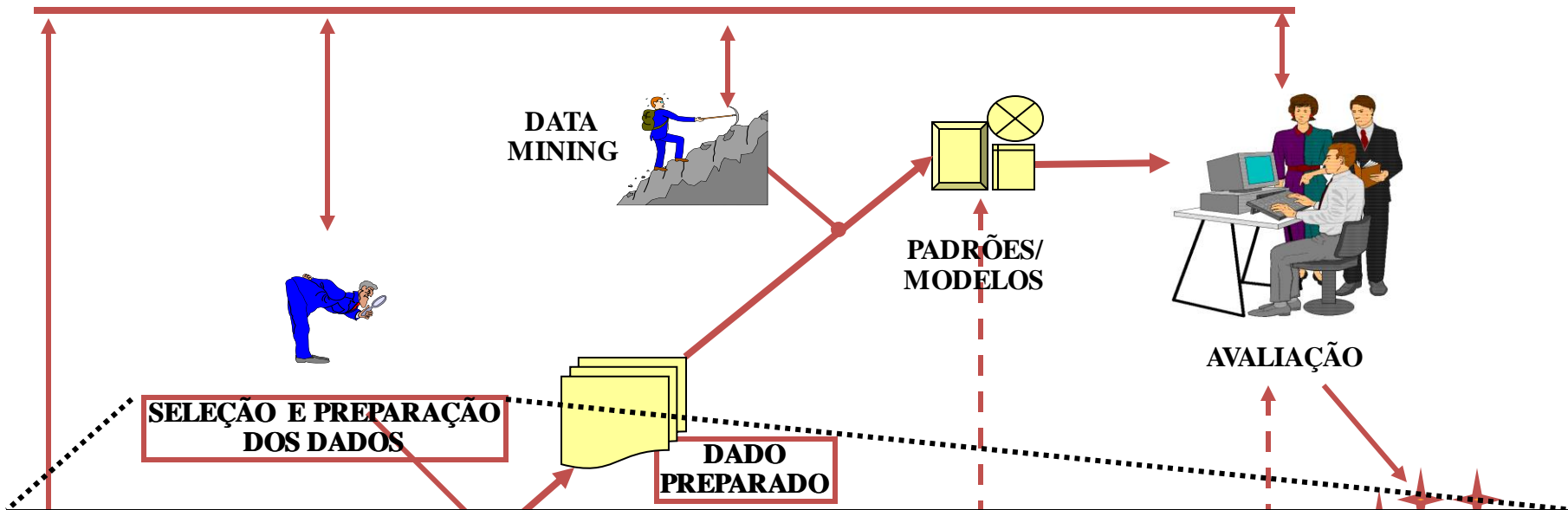
Etapas do Processo KDD



Etapas do Processo KDD



Etapas do Processo KDD

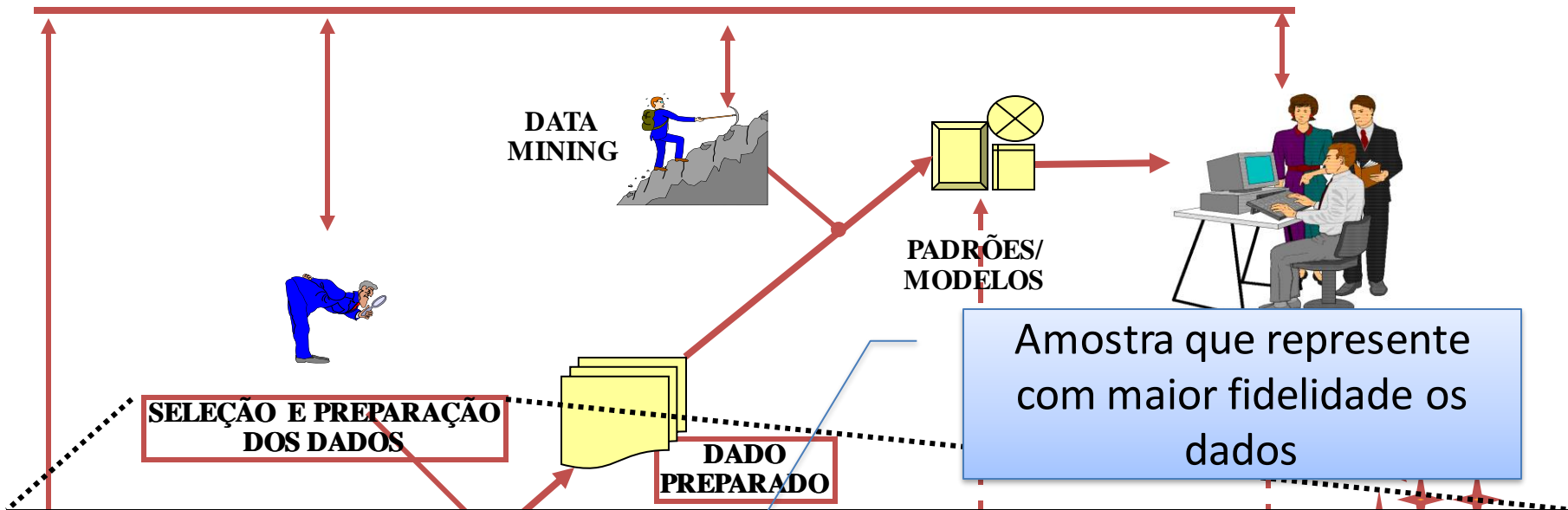


Alguns problemas da extração de conhecimento a partir de grandes dados:

- 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

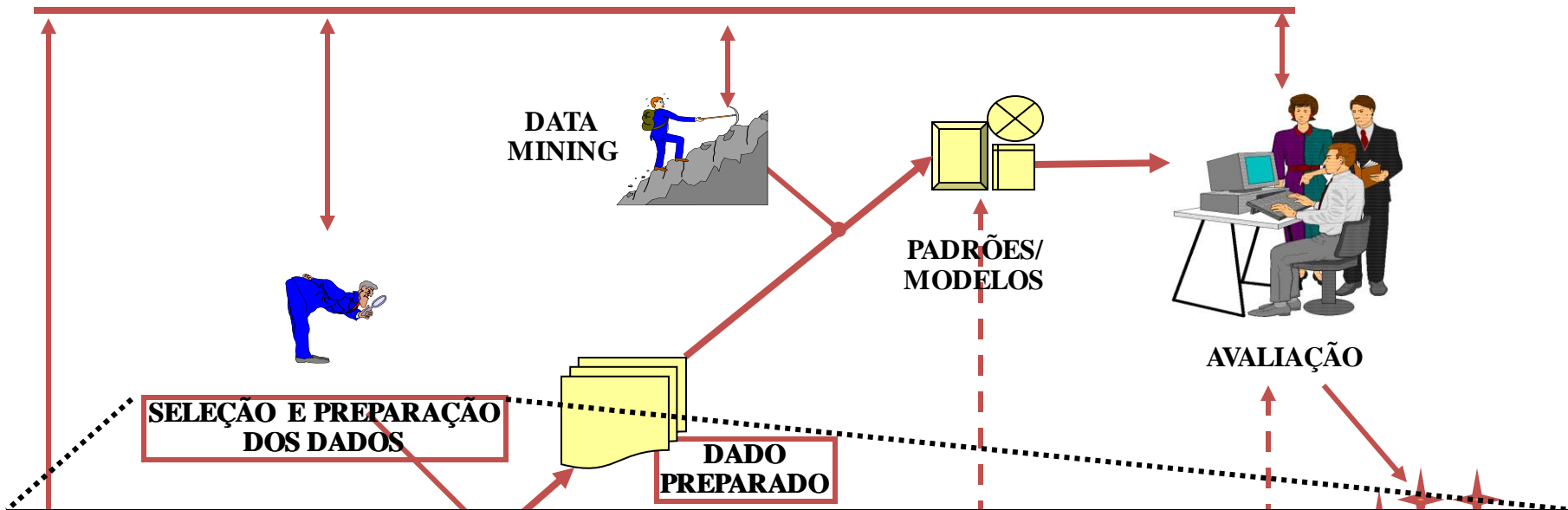
Etapas do Processo KDD



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

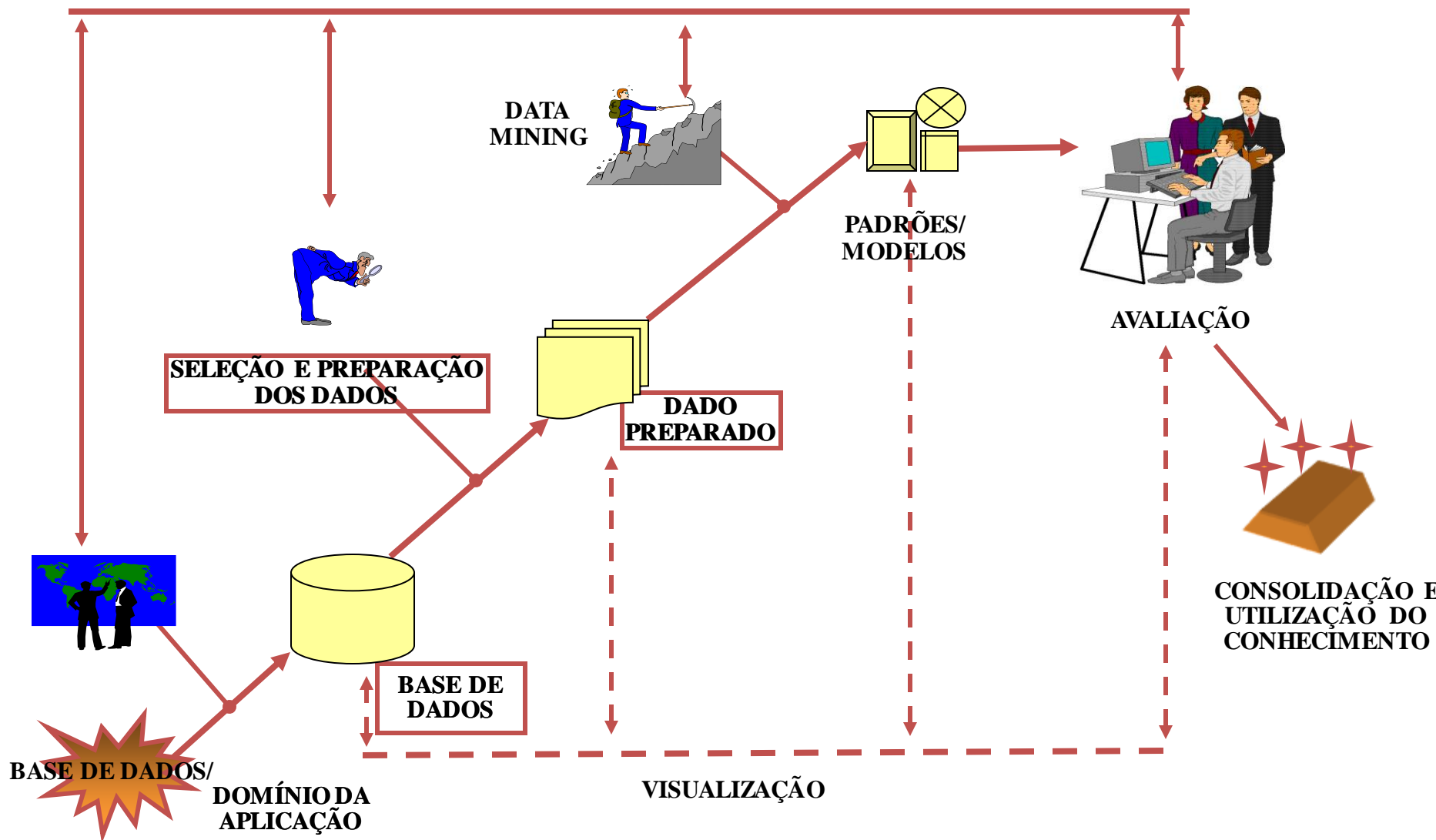
Etapas do Processo KDD



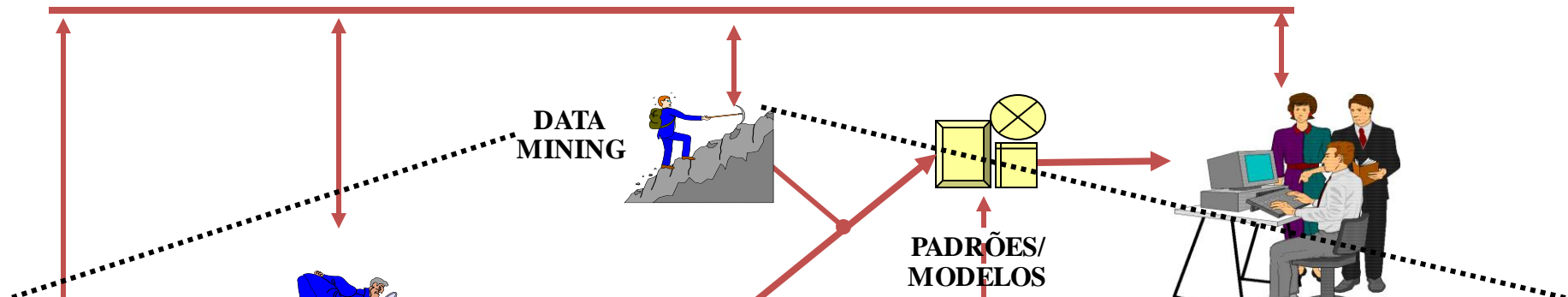
A preparação e redução da amostra envolve a observação dos seguintes aspectos:

- Eliminação dos registros duplicados, lixo nos dados.
- Tratamento de ruídos nos dados
- Manipulação de valores de atributos ausentes
- Encontrar métodos para reduzir efetivamente o número de variáveis a serem consideradas no processo

Etapas do Processo KDD



Etapas do Processo KDD

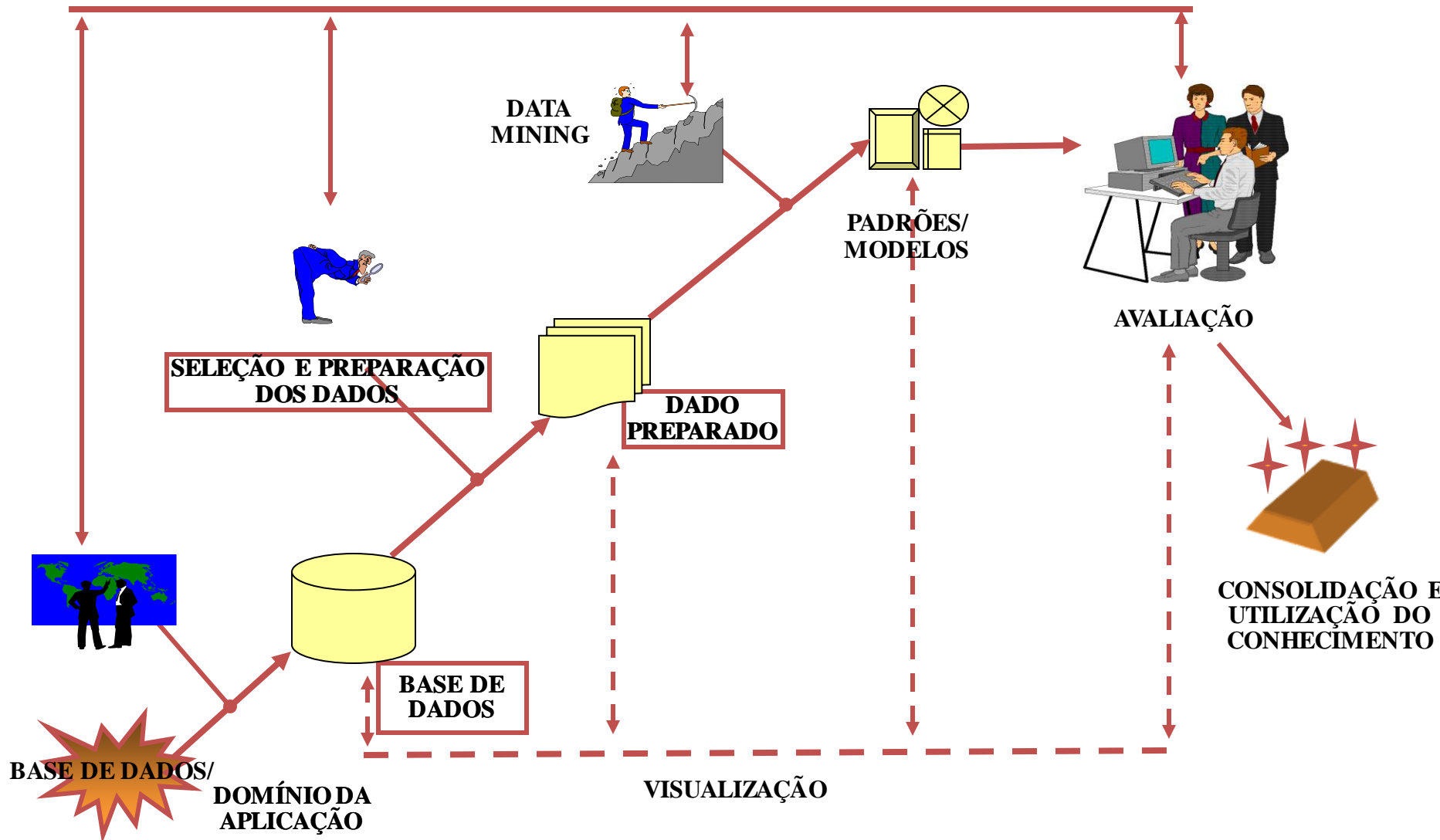


Data Mining (DM) ou Mineração de Dados (MD) envolve a utilização de algoritmos para extração de padrões válidos, compreensíveis e potencialmente úteis nos dados.

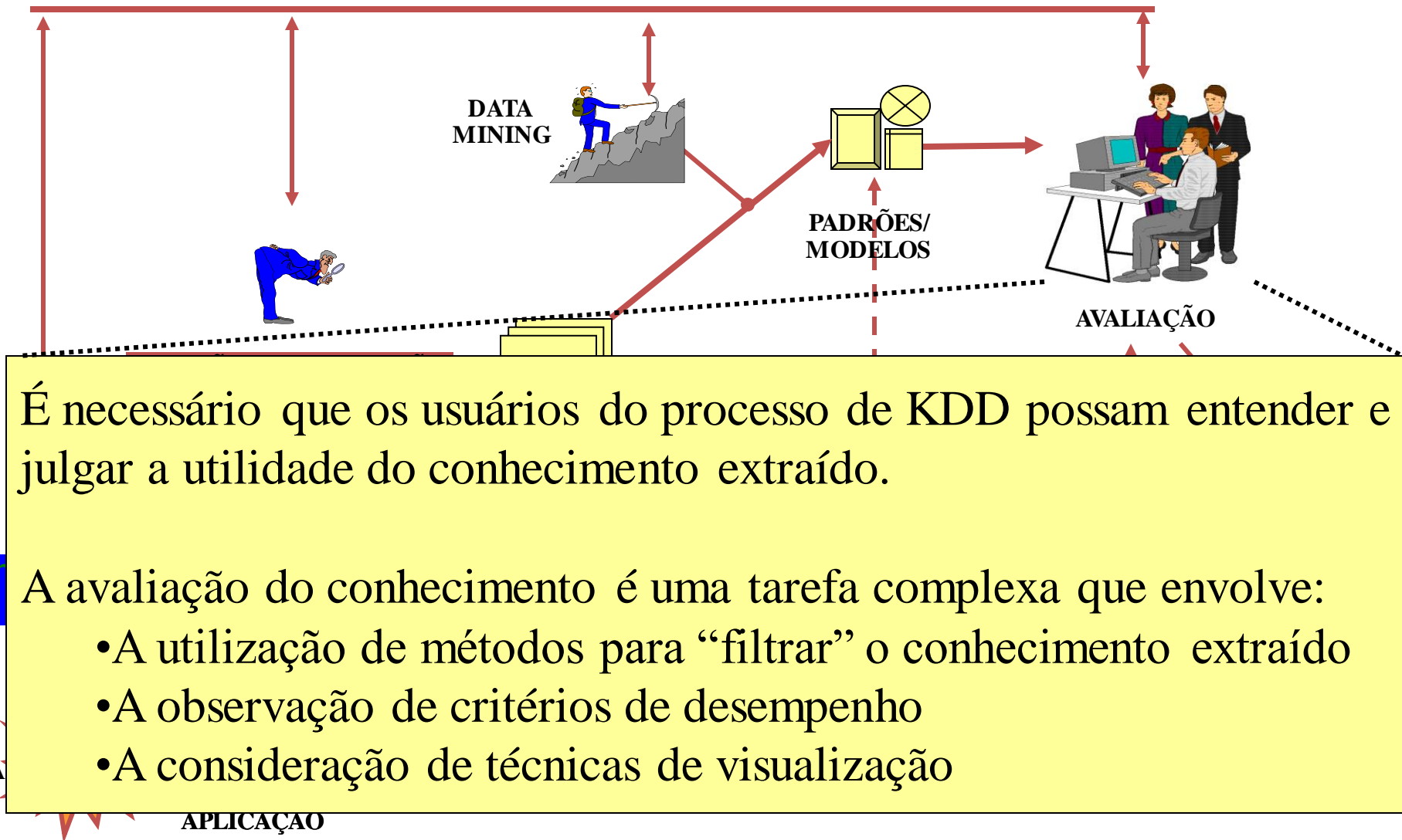
Esses algoritmos consistem da combinação de três componentes:

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

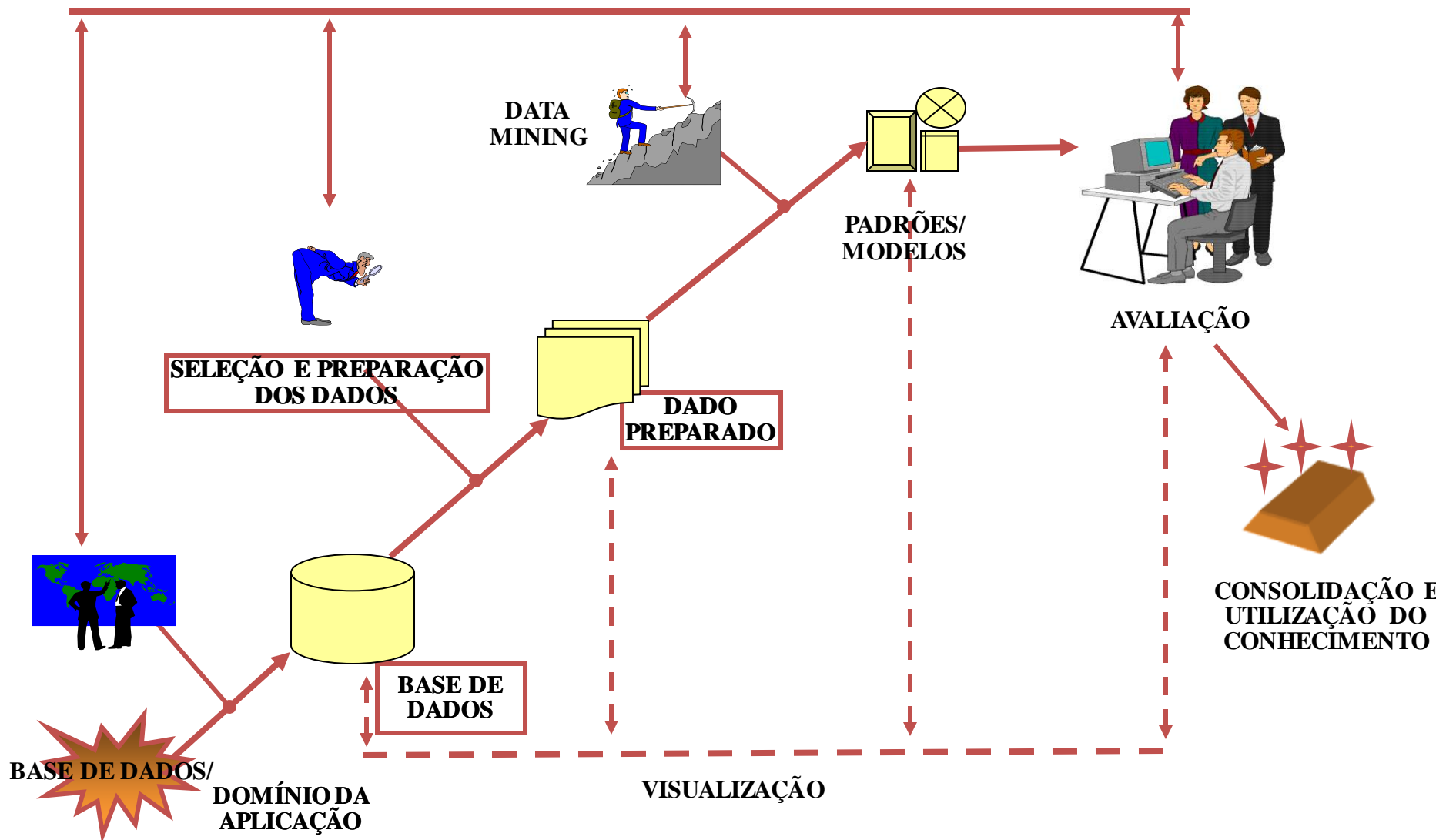
Etapas do Processo KDD



Etapas do Processo KDD



Etapas do Processo KDD

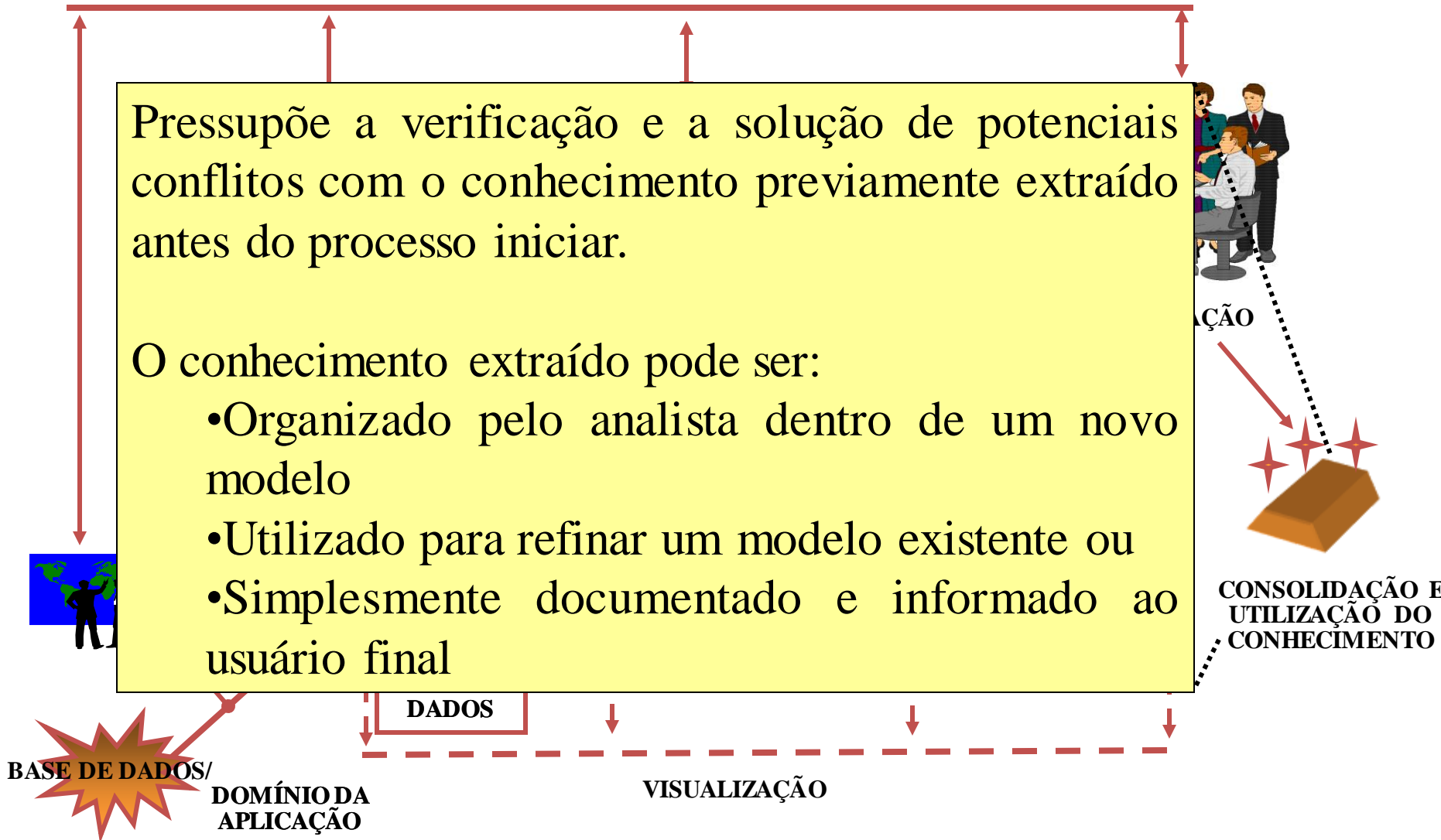


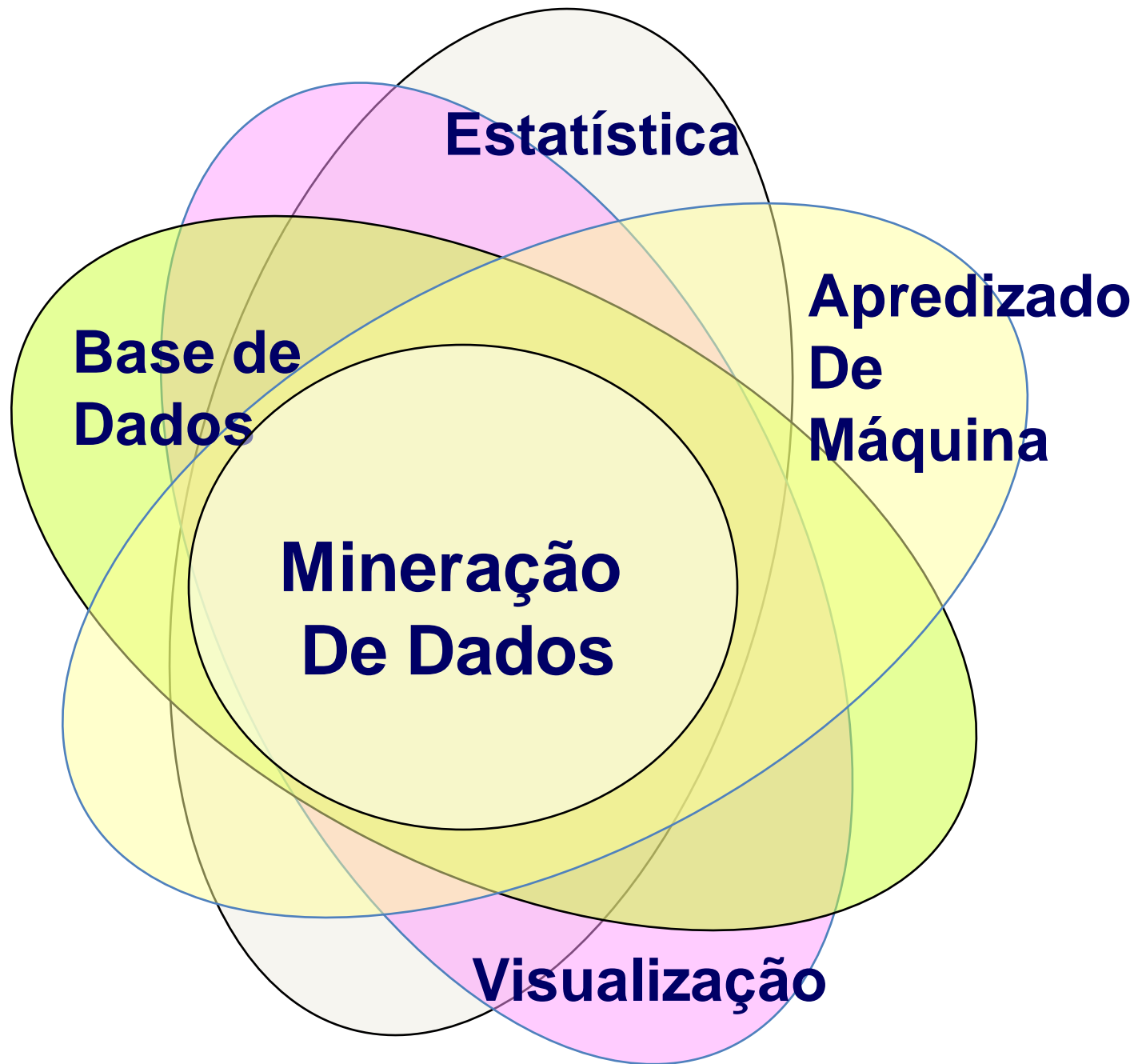
Etapas do Processo KDD

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

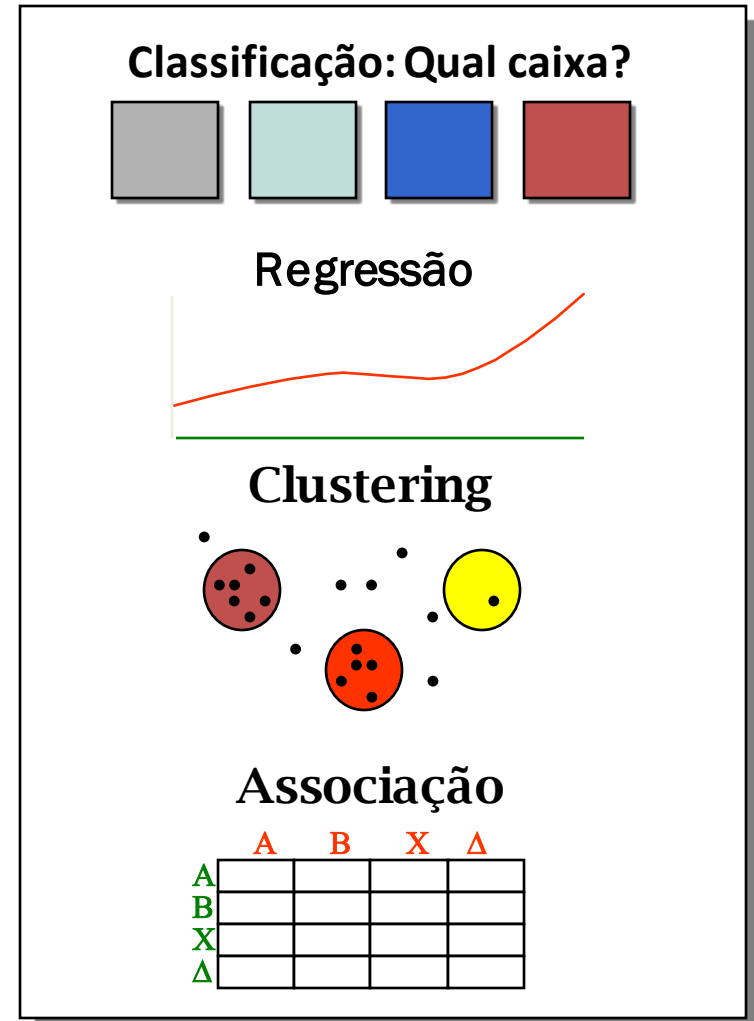




Tarefas em Mineração de Dados

(focadas em Aprendizado de Máquina)

- Predição:
 - Classificação
 - Regressão
- Clustering
- Associação



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

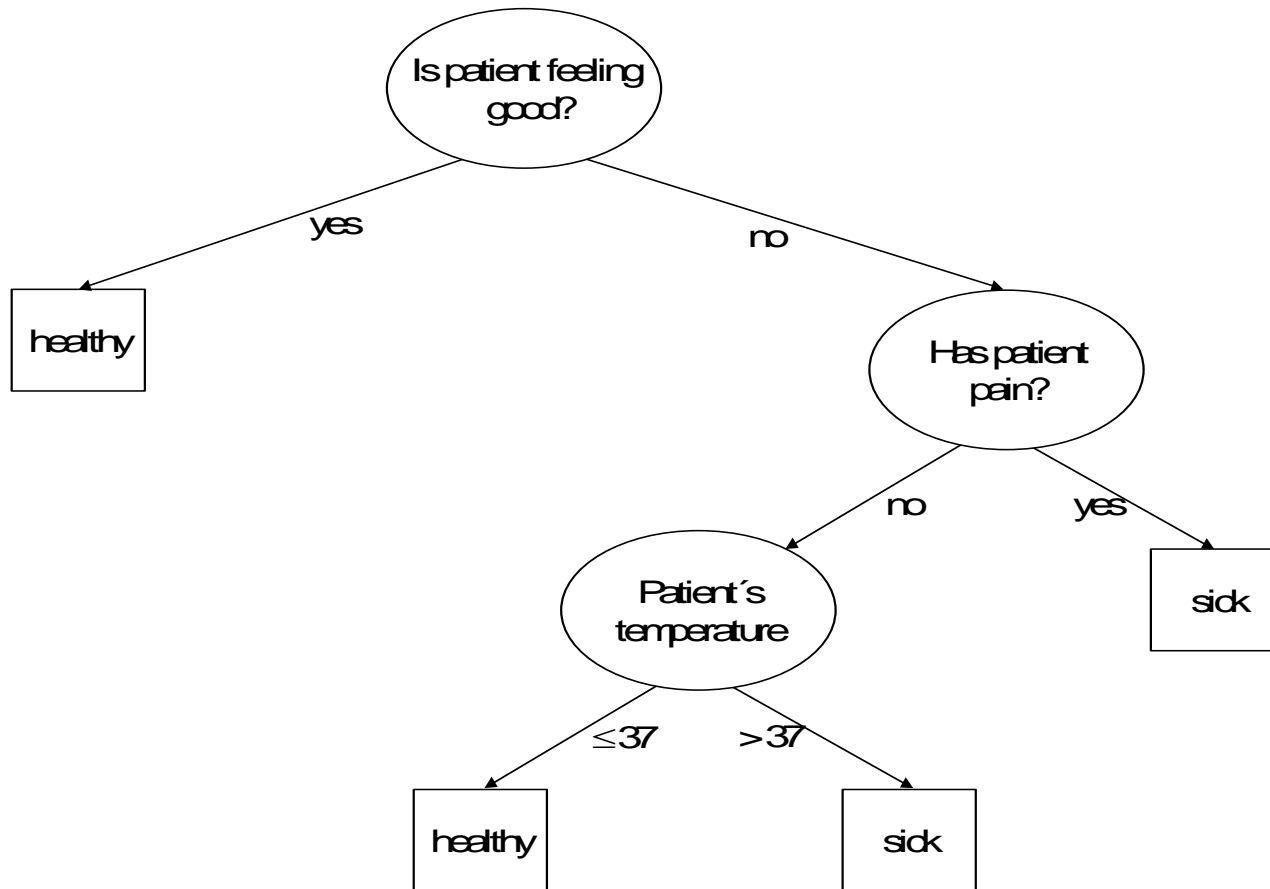


Classificação

- 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

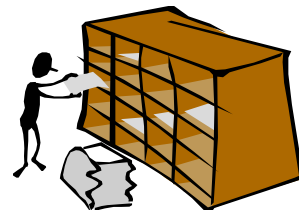


Classificação



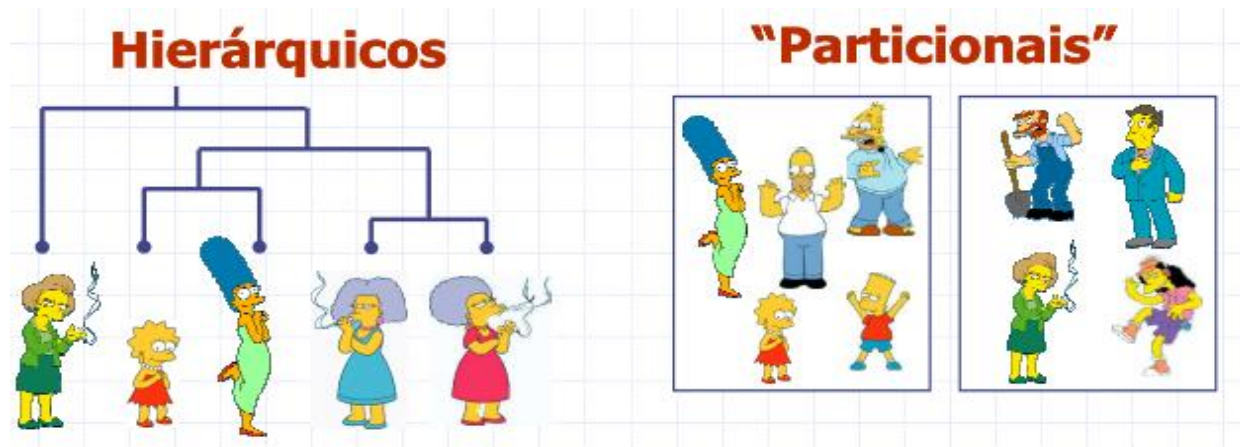
Clustering

- **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



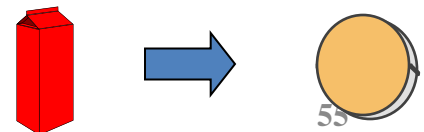
Associação

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

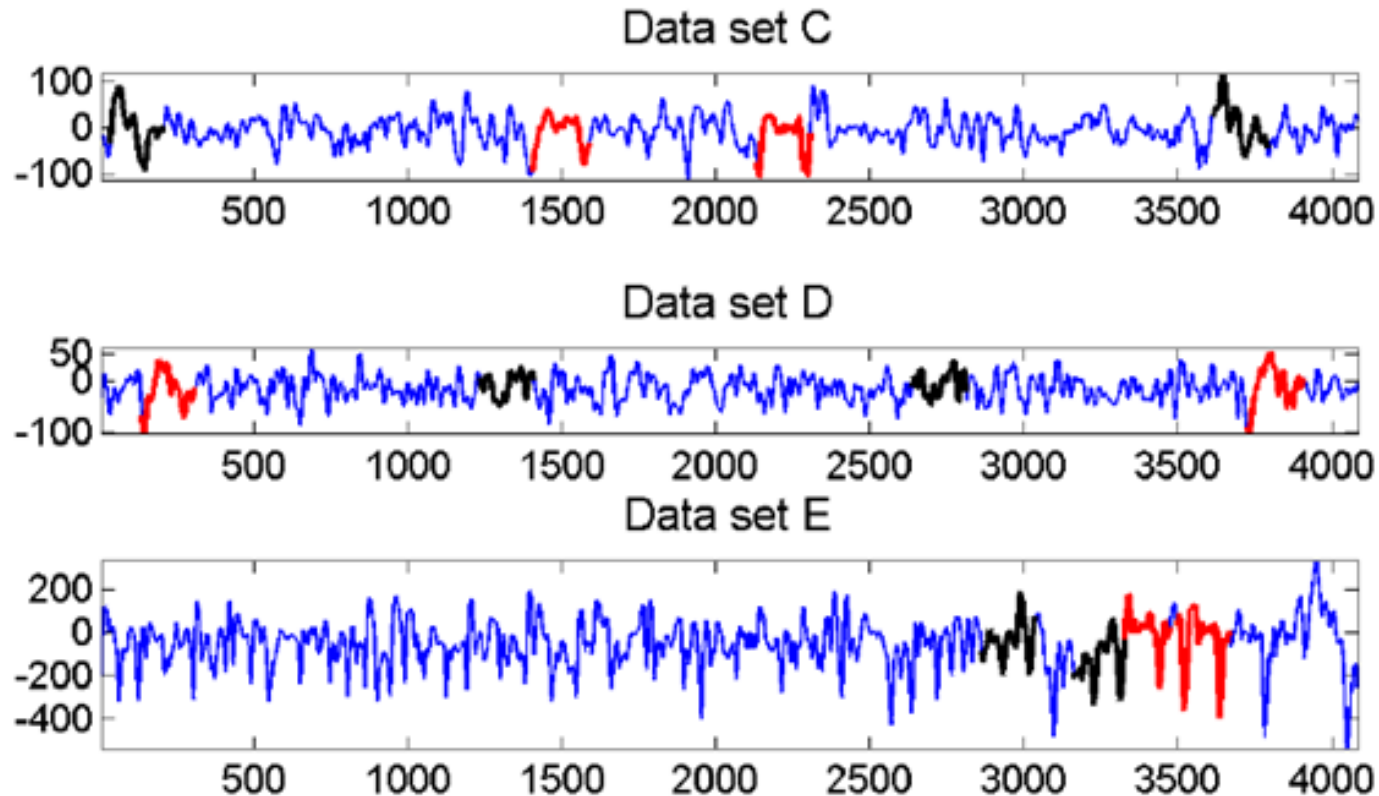
$$a_1 \wedge a_2 \wedge \dots \wedge a_n \rightarrow b_1 \wedge b_2 \wedge \dots \wedge b_n$$

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

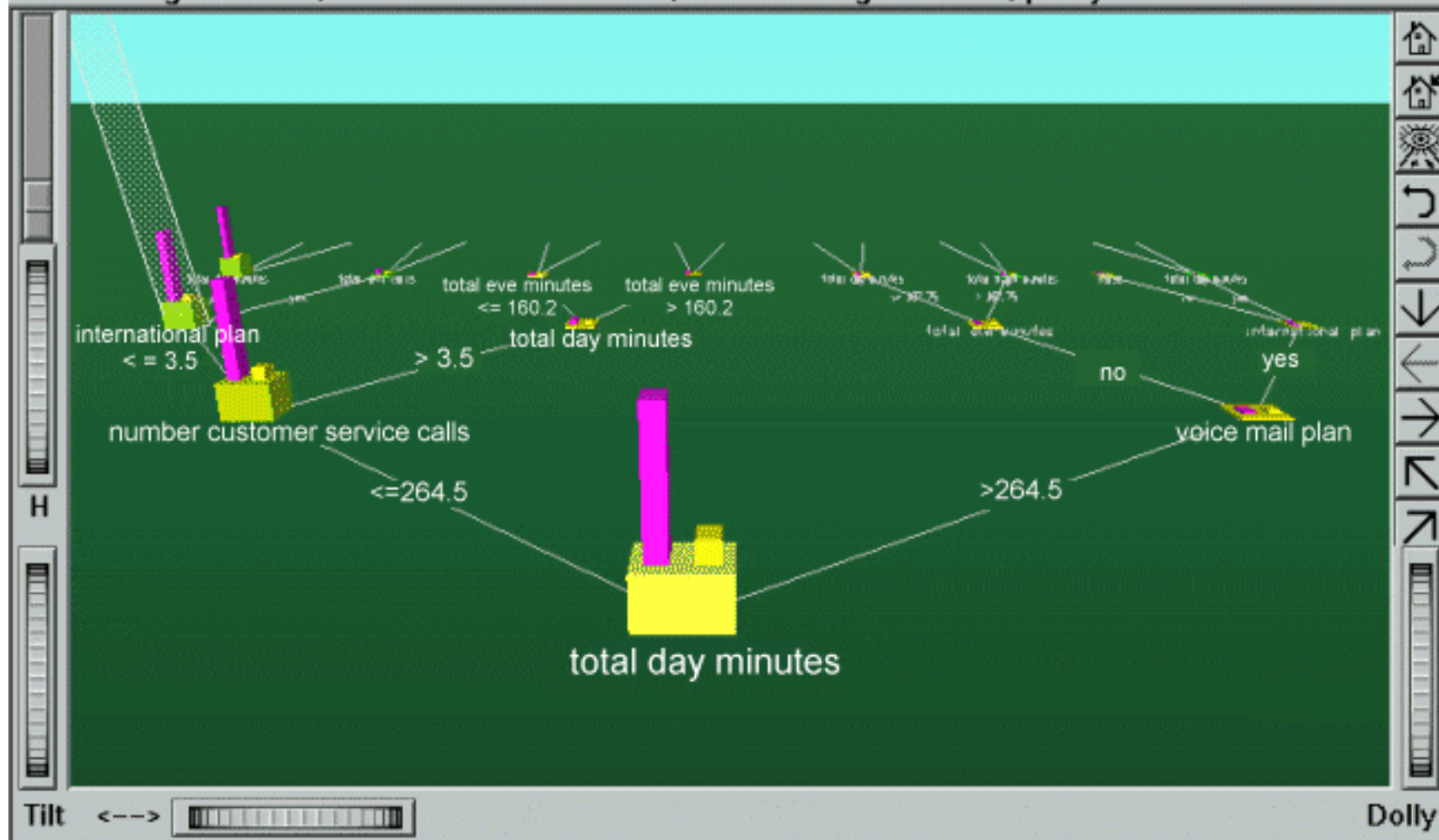
Ferramentas

- 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

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

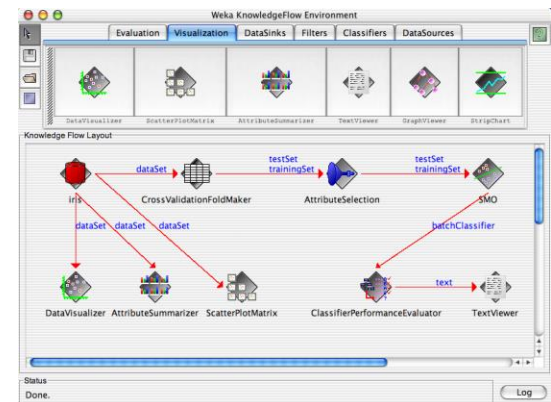
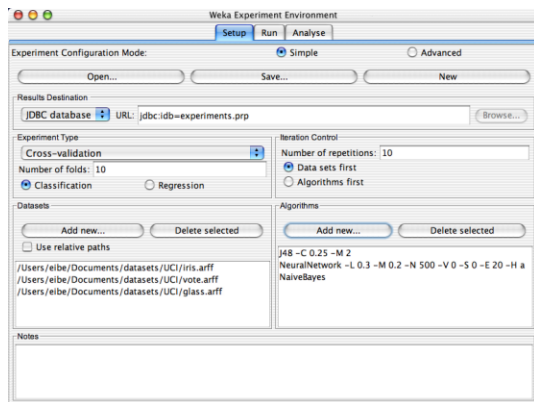
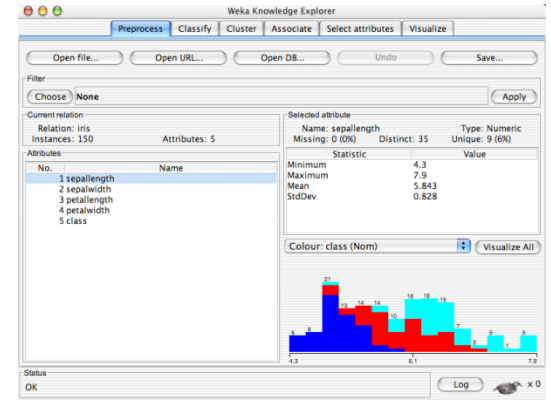
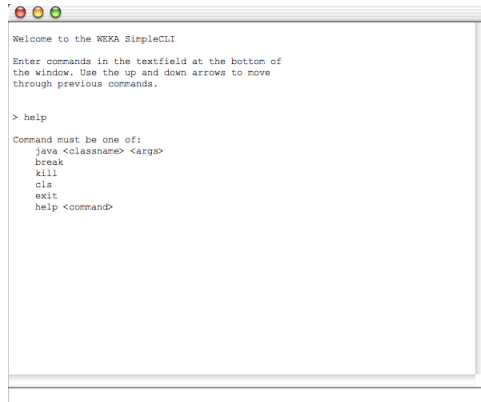
Subtree weight:5000.00, test-set error:5.46+/-0.56, test-set weight:1667.00, purity:41.21



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