

Inteligência Artificial

Mineração de Dados e/ou Ciência de Dados

Prof. Fabio Augusto Faria

1º semestre 2021



Agenda

- Introdução
- Mineração de Dados x Ciência de Dados
- Etapas da Mineração de Dados
- Aplicações Reais
- Pesquisas no IC-UNIFESP
- Conclusões

Introdução

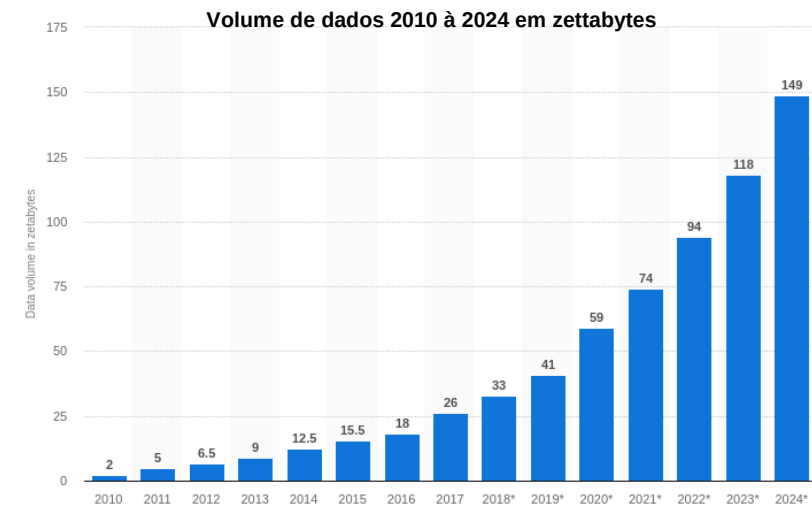
- Avanço das tecnologias de aquisição e armazenamento de dados
- Acesso à internet:
 - ~3,7 Bilhões de pessoas
 - 3,5 Bilhões de buscas no Google por dia
 - dispositivos móveis (e.g., celular e tablet)
- Uso das mídias sociais por minuto:
 - Usuários do Snapchat compartilham 527.760 fotos
 - 4.146.600 vídeos assistidos no Youtube
 - 456.000 tweets no Twitter
 - 46.740 fotos postados no Instagram

Introdução

- “The Data Revolution”:
 - 2.5 quintilhões de bytes de dados por dia
 - 1.7MB de dados foram criados por segundo em 2020
 - ~90% de todos os dados foram criados nos últimos 2 anos
 - Dados não-estruturados é o problema para 95% das empresas
 - Em 2023, a indústria do “BIG DATA” terá um valor estimado de US \$77 bilhões
 - BIG DATA é dependente de sistemas automáticos até 2020

- Fotos:

- 1.2 Trilhão de fotos tiradas até 2017
- 4.7 trilhão de fotos armazenadas



Introdução

- Necessidade de novas ferramentas:
 - Tratamento de dados
 - Análise de dados
 - Exploração de dados
 - Descobrimento de conhecimento
- Mineração de Dados e/ou Ciência de Dados

Introdução

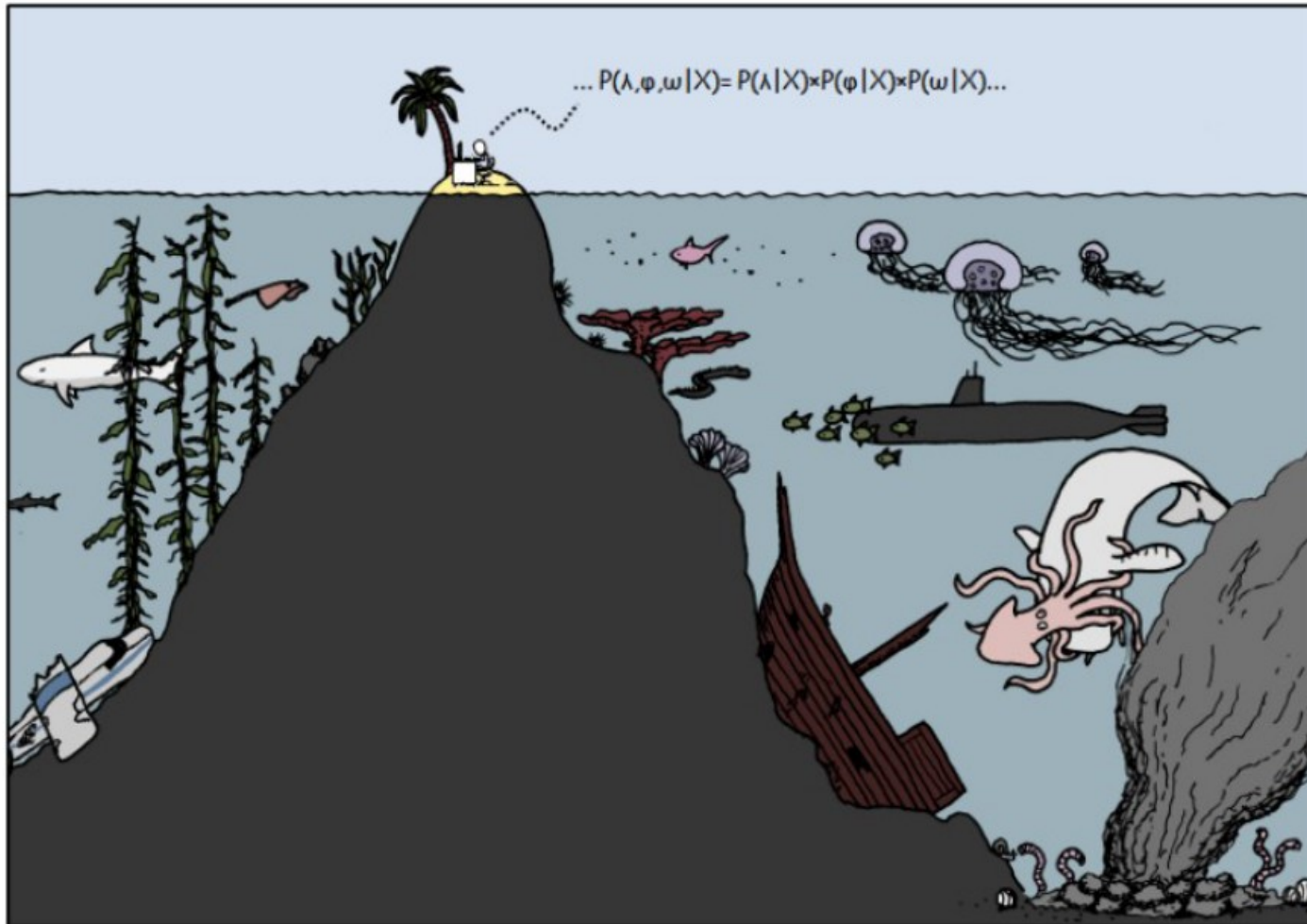
What is “reasoning for complex data” ? Well, let me explain in two pictures:



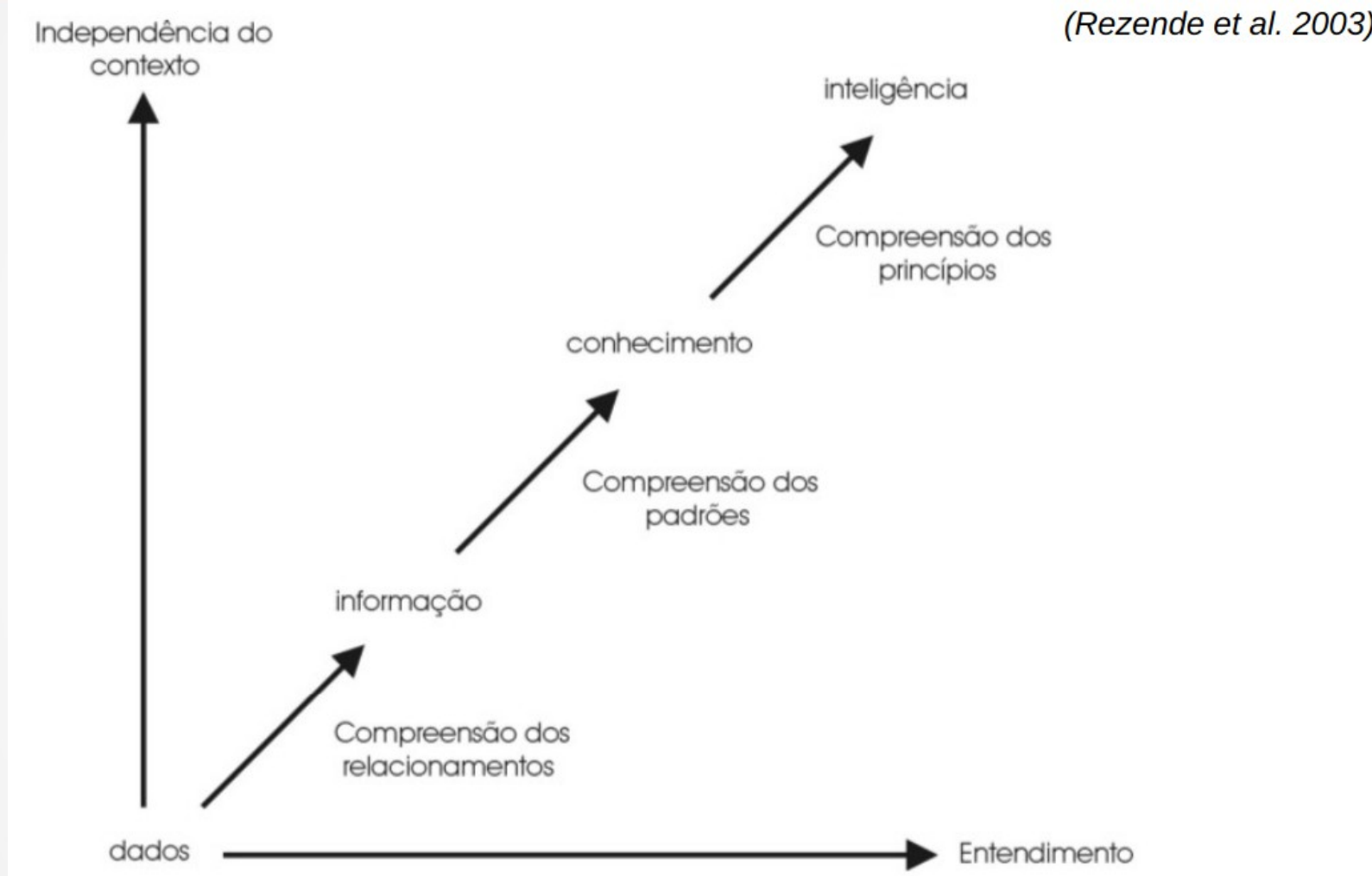
This is when you look at data, but don't reason on it...

Introdução

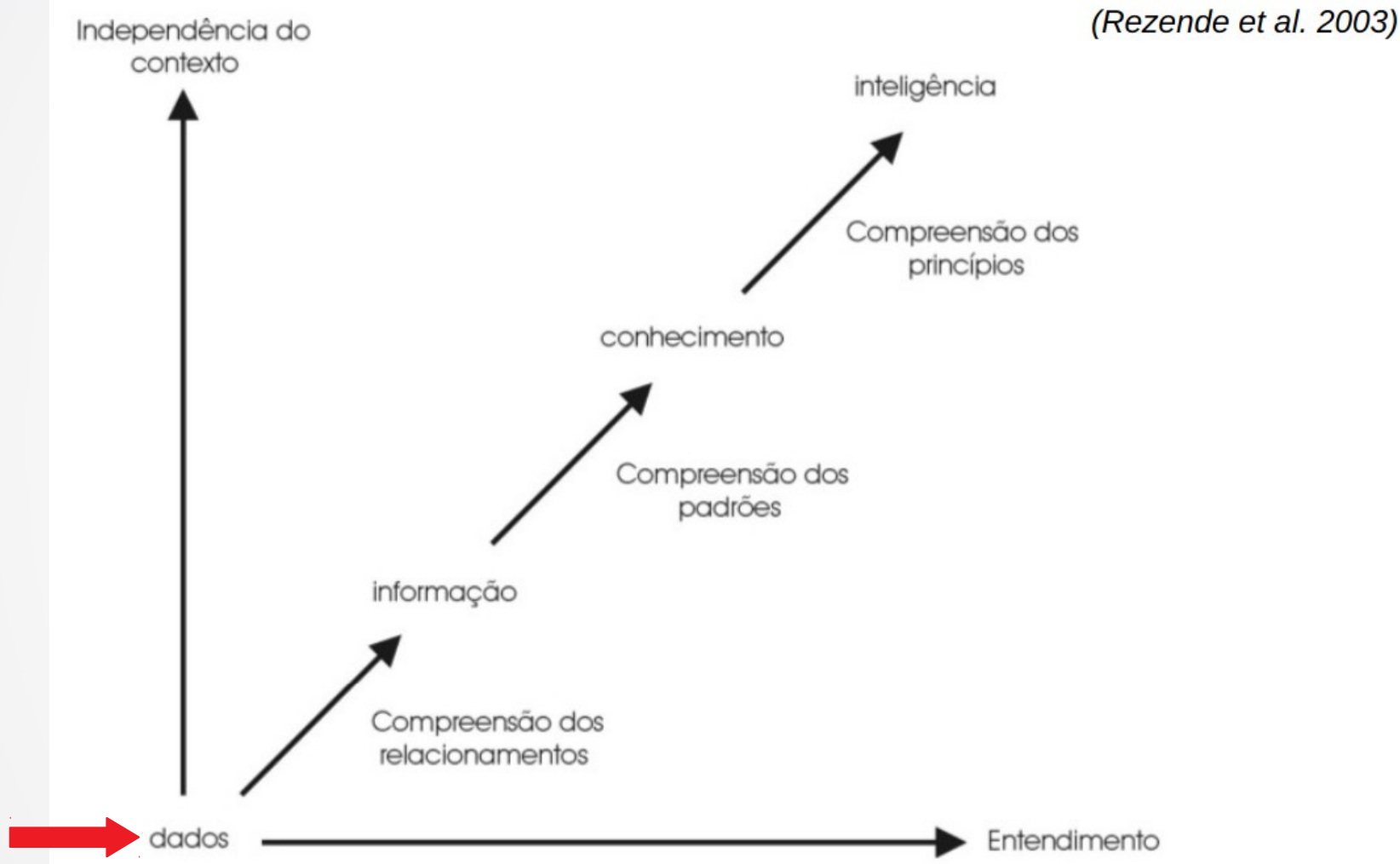
...and this is when reasoning techniques allow you understand it.
Got the difference ?!



Dados, Informação e Conhecimento

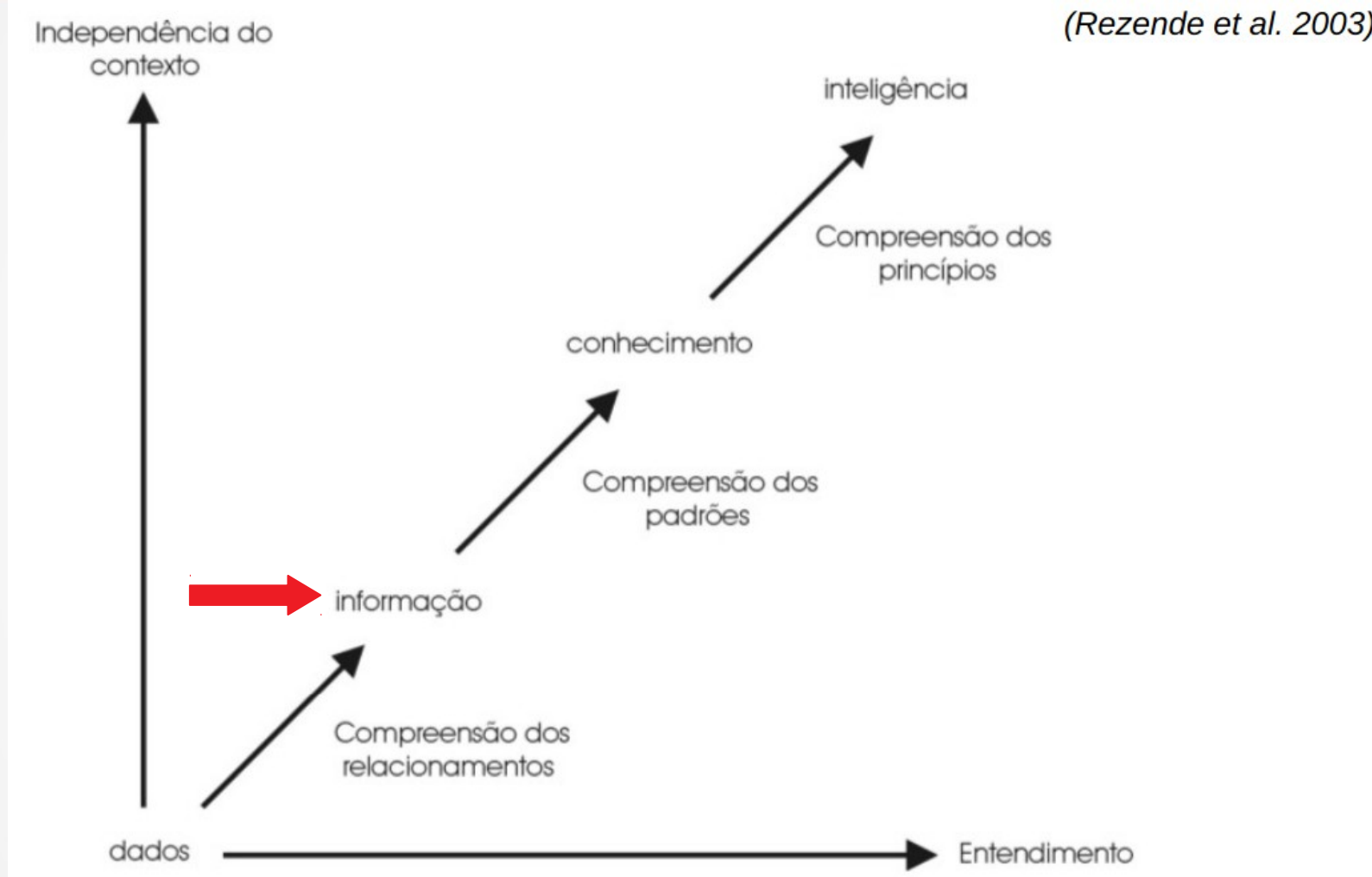


Dados, Informação e Conhecimento



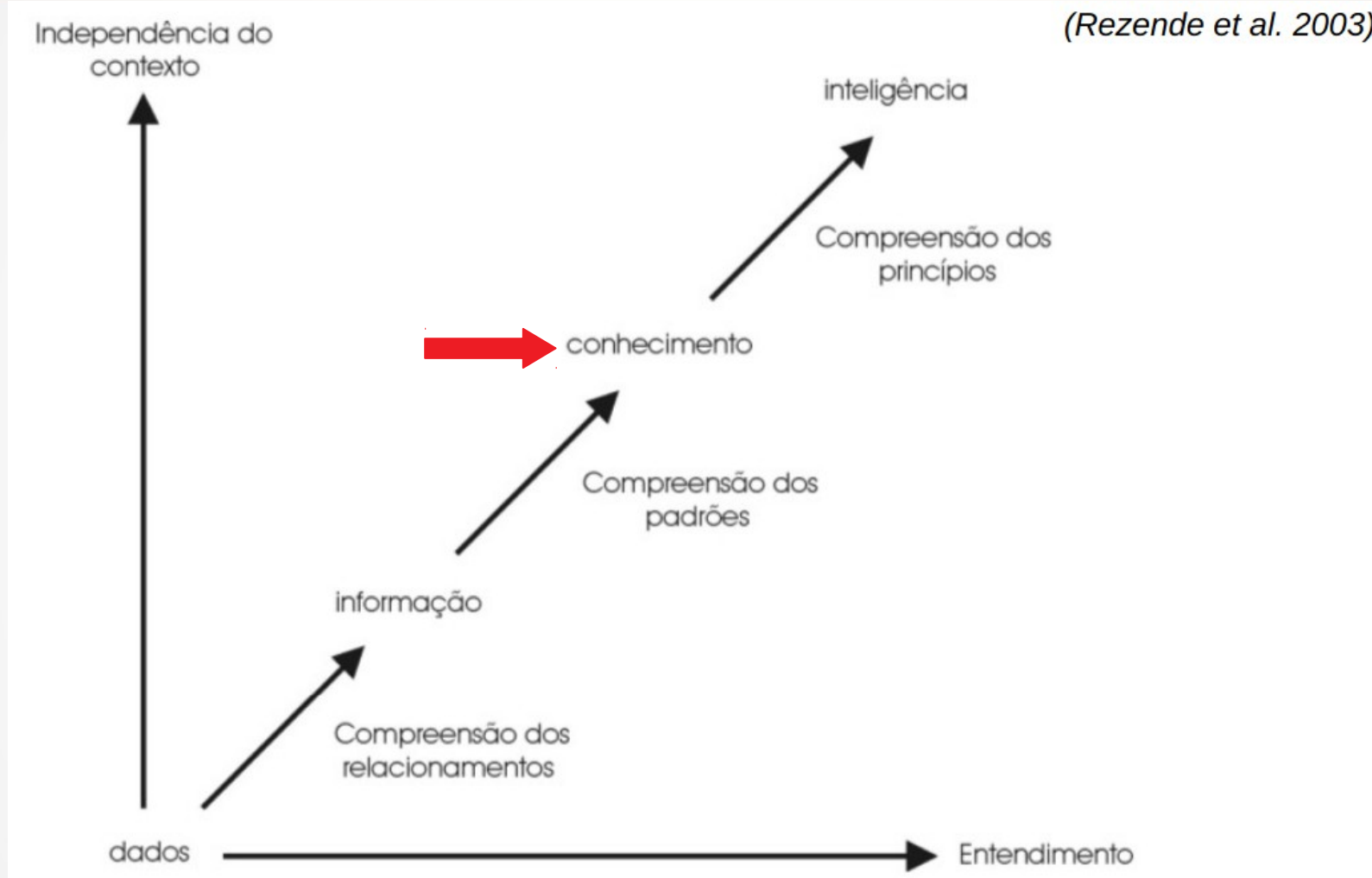
“Elemento puro, quantificável sobre um determinado evento (e.g., fatos, números, texto ou multimídia) que pode ser processado.”

Dados, Informação e Conhecimento



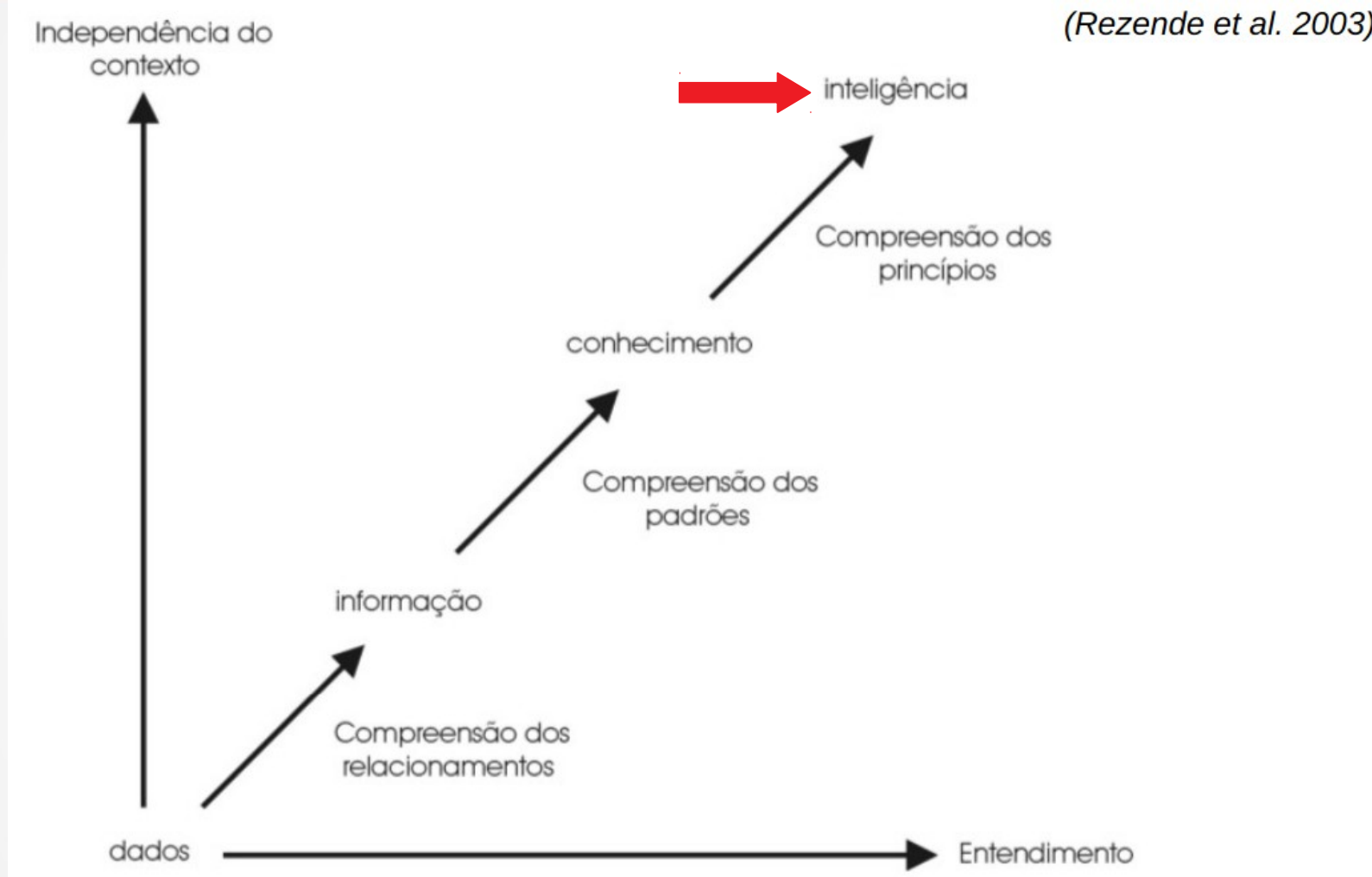
“Envolve a interpretação de um conjunto de dados, considerando padrões, associações Ou relações que todos aqueles dados acumulados podem proporcionar.”

Dados, Informação e Conhecimento



“Informação pode gerar conhecimento que ajude na análise de padrões históricos para conseguir uma previsão dos fatos futuros.”

Dados, Informação e Conhecimento



“A compreensão, análise e síntese, necessárias para a tomada de decisões inteligentes, são realizadas a partir do nível do conhecimento.”

Mineração x Ciência de Dados

Qual a diferença?

Mineração x Ciência de Dados

No mercado de trabalho ([LinkedIn](#)):

Mineração de Dados $=$ Ciência de Dados

Mineração de Dados \neq Ciência de Dados

Mineração de Dados \simeq Ciência de Dados

Requirements

What we are looking for:

- Deep understanding of machine learning concepts: regression and classification, clustering, neural networks, feature selection, cross-validation, curse of dimensionality, bias-variance tradeoff, model explainability, etc.;
- Strong knowledge of probability and statistics, including experimental design, predictive modeling, optimization, and causal inference;
- Good understanding of the engineering challenges to deploy machine learning systems to production;
- Proficiency in Python or another major programming language;
- Someone up-to-date with recent advances in Machine Learning, and willing to share his/her knowledge with the other members of the Data Science Chapter;
- Excellent written and verbal technical communication skills;
- Experience leading teams and managing careers.

Mineração x Ciência de Dados

No mercado de trabalho ([LinkedIn](#)):

Mineração de Dados $=$ Ciência de Dados

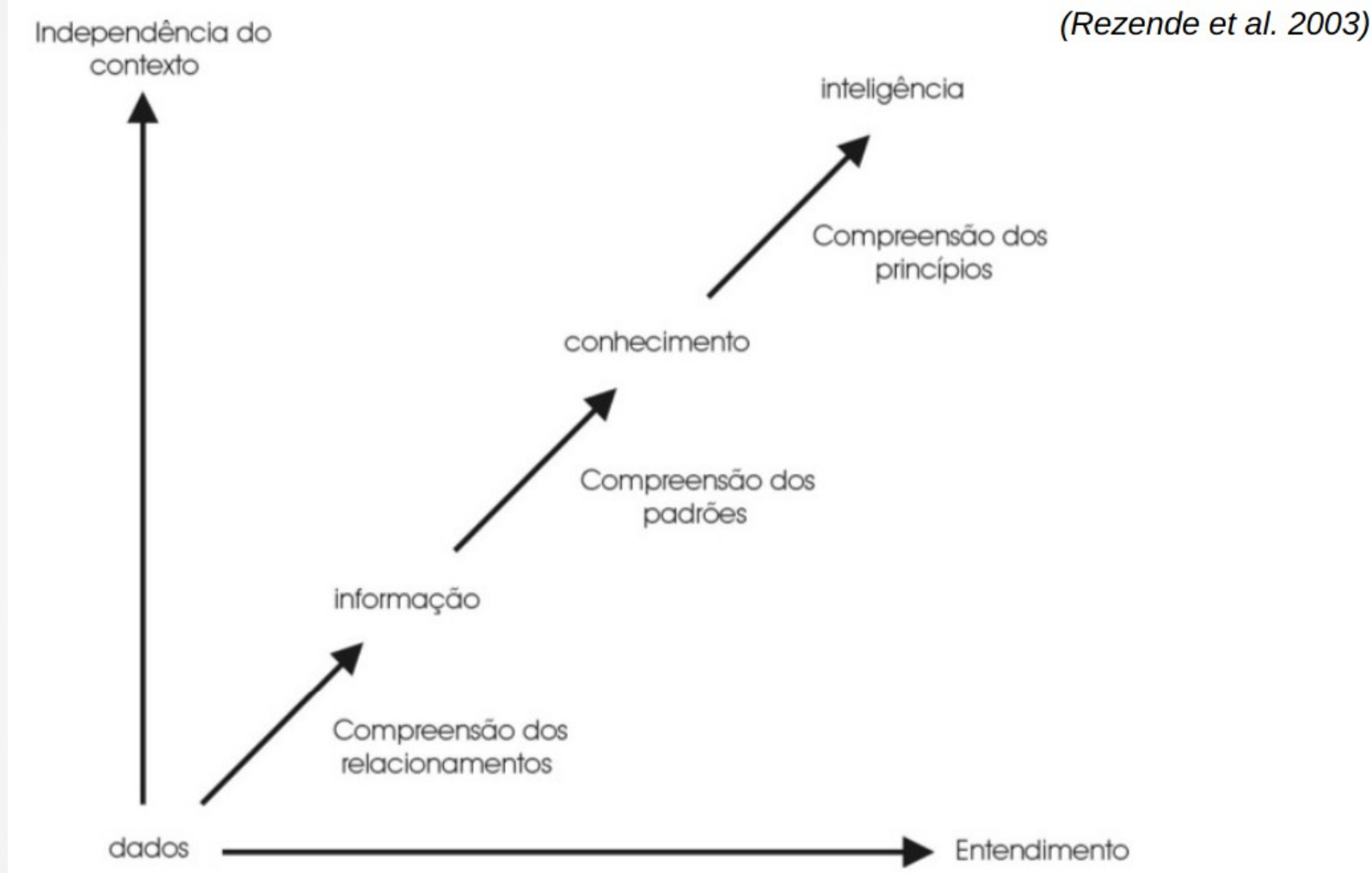
Mineração de Dados \neq Ciência de Dados

Mineração de Dados \approx Ciência de Dados

“As rotinas de Ciência de dados pode fazer parte do processo de Mineração de Dados da aplicação alvo.”

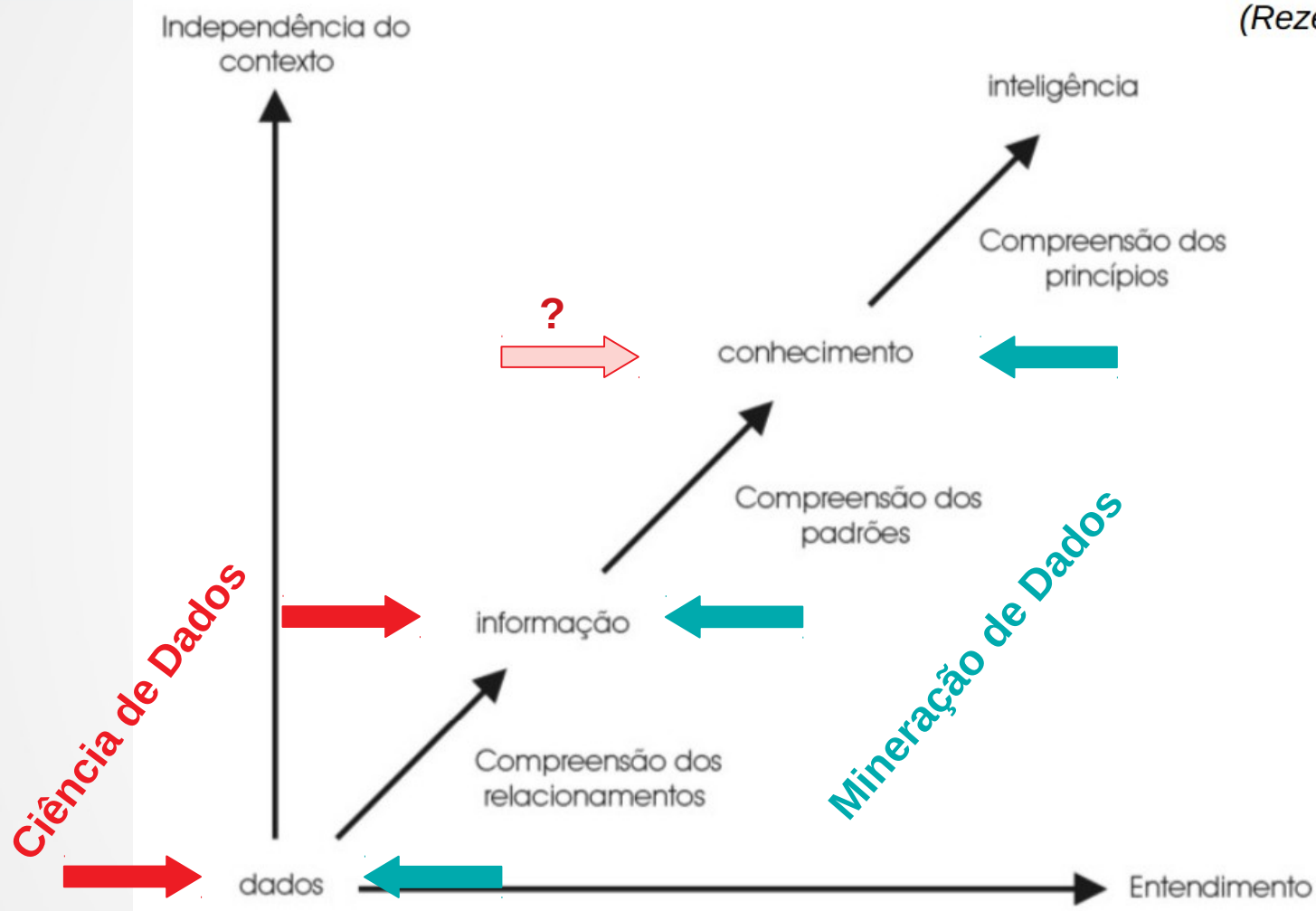
Fabio Faria

Mineração x Ciência de Dados



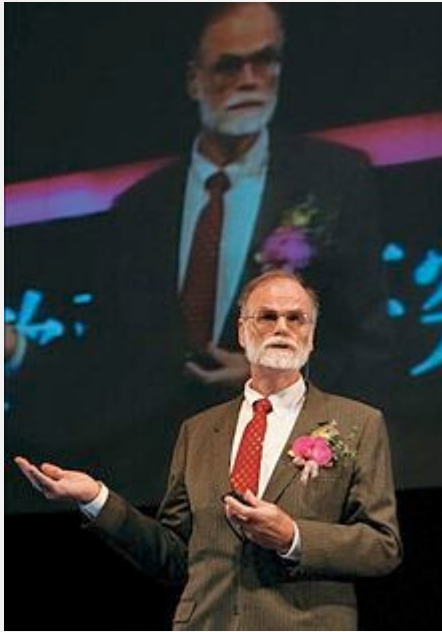
Mineração x Ciência de Dados

(Rezende et al. 2003)



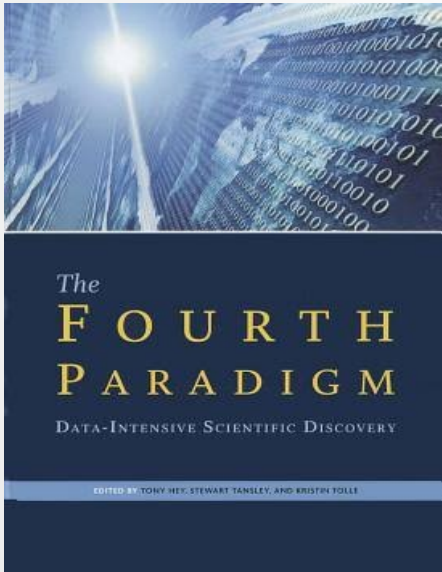
***ESTA É MINHA INTERPRETAÇÃO SOBRE OS PROFISSIONAIS!**

Ciência de Dados



Segundo **Jim Gray** (prêmio Turing 1998), a ciência de dados como um "quarto paradigma" da ciência (empírica, teórica, computacional e agora baseada em dados) e afirmou:

"Tudo na ciência está mudando por causa do impacto da tecnologia da informação e do dilúvio de dados."



In *The Fourth Paradigm: Data-Intensive Scientific Discovery*, the collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados.”

Rezende et al. 2003

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos **dados**.”

Rezende et al. 2003

Dados: repositório de dados do domínio da aplicação alvo que serão analisados.

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o processo de identificação de **padrões** válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados.”

Rezende et al. 2003

Padrões: Denota alguma abstração de um subconjunto dos dados em alguma linguagem descritiva de conceitos.

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o **processo** de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados.”

Rezende et al. 2003

Processo: Uma atividade que envolve diversas etapas.

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o processo de identificação de padrões **válidos**, novos, potencialmente úteis e compreensíveis embutidos nos dados.”

Rezende et al. 2003

Válidos: padrões descobertos devem possuir algum grau de certeza (validade).

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, **novos**, potencialmente úteis e compreensíveis embutidos nos dados.”

Rezende et al. 2003

Novos: padrão encontrado deve fornecer novidades sobre os dados.

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente **úteis** e compreensíveis embutidos nos dados.”

Rezende et al. 2003

Úteis: padrões descobertos devem ser utilizado.

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de Conhecimento de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e **compreensíveis** embutidos nos dados.”

Rezende et al. 2003

Compreensível: usuários devem entender os padrões descobertos e poder analisá-los mais a fundo

Mineração de Dados

- Segundo Fayyad et al. 1996:

“Extração de **Conhecimento** de Base de Dados é o processo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis embutidos nos dados.”

Rezende et al. 2003

Conhecimento: relacionados com medidas de utilidade, originalidade e compreensão do domínio aplicado (resultado final).

Mineração de Dados

- Multidisciplinar:
 - Banco de Dados
 - Estatística
 - Inteligência Artificial
 - Aquisição de Conhecimento
 - Visualização de Dados
 - Computação de Alto Desempenho
 - Processamento de Dados Multimídia

Mineração de Dados

- Descoberta de Conhecimento x **Mineração de Dados**



Mineração de Dados

- Descoberta de Conhecimento == Mineração de Dados



Conhecimento do Domínio



Conhecimento do Domínio

- Identificação do problema
- Definição de objetivos e metas

Participação dos especialistas é essencial em TODAS as fases!

Pré-processamento



Pré-processamento

- Formatação dos dados (e.g., entrada de algum algoritmo)
- Redução do volume
- Transformação de dados (e.g., tipos dos dados)
- Seleção de dados (e.g., filtros)
- Extração dos dados (e.g., descritores de imagens)

Identificação da representatividade das amostras para o problema.

Os dados são suficientes para representar o Mundo Real?

Extração de Padrões



Extração de Padrões

- Escolher a tarefa (e.g., **classificação**, regressão e regras de associação)
- Escolher o algoritmo (e.g., **kNN**, **SVM**, **DT**, e **Naive Bayes**)
- Executar a extração de padrões

Extração de Padrões

- Escolher a tarefa
 - De acordo com os objetivo (e.g., **dados bancários**)
- Tipo de tarefas
 - Classificação: **rotular** cliente em BOM ou MAU PAGADOR
 - Regressão: **estimar** um grau (**valor contínuo**) de confiabilidade do cliente
 - Regras de Associação: **associar** movimentação do cliente
 - Se CASA > \$500mil & CARRO > \$200mil
 - Logo, RENDA > \$30mil

Extração de Padrões

- Escolher o algoritmo
 - De acordo com a tarefa escolhida
 - Complexidade do problema
 - Não existe uma “bala de prata”
 - Buscar na literatura
 - Testar diferentes técnicas
 - Analisar os resultados na etapa de pós-processamento
- Extração de padrões
 - Emprego dos algoritmos escolhidos nos dados da aplicação alvo

Etapas da Mineração de Dados



Pós-processamento

- Análise dos padrões descobertos
- Descoberta de possíveis soluções para o problema alvo
- Avaliação de conhecimento (e.g., desempenho e qualidade)

1- Se existirem conhecimentos NOVOS e ÚTEIS então poderão ser utilizados para alguma tomada de decisão;

Ou

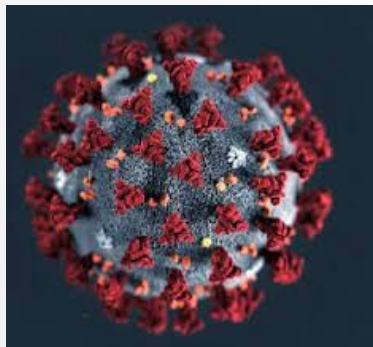
2- Caso contrário, repete-se parte ou todo processo com diferentes ajustes.

Dados Complexos

- Textuais
- Multimídia (imagem, vídeo e som)
- Séries temporais
- Geográficos
- Heterogêneos
- Outros.

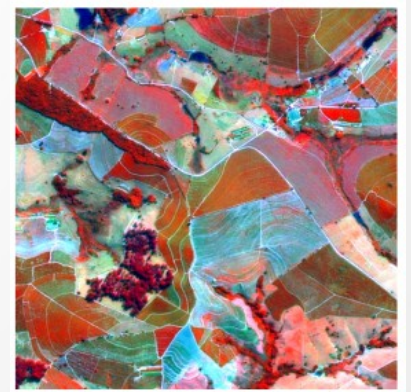
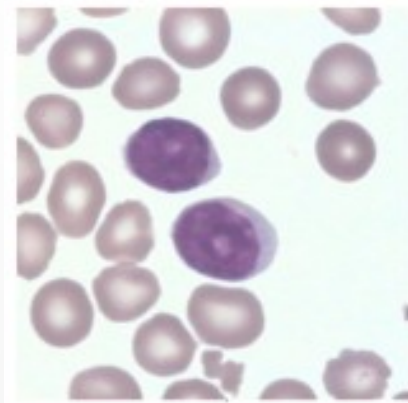
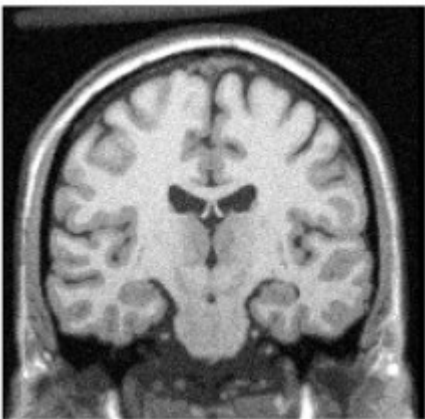
Aplicações

- Serviço de Recomendação (e.g., Amazon, Netflix, Nike e Walmart)
- Análise de Sentimento (e.g., twitter, orkut e facebook)
- Detecção de Anomalias (e.g., fraudes bancárias)
- Sistema de Previsões (e.g., desastres, mercado financeiro e atentados)
- Identificação de Perfis (e.g., bancos e financeiras)
- Sistema Biométrico (e.g., face, digitais, iris e voz)
- Entender melhor o “desconhecido” (COVID-19)



Minhas pesquisas

- Auxílio ao Diagnóstico Médico (MRI cérebro)
- Identificação de Pragas de Plantação (moscas-de-frutas)
- Reconhecimento de Regiões de Plantação e Florestas (café e Amazônia)
- Sistema Forense (*splicing e composition*)
- Monitoramento de Vegetação (aquecimento global)



Bibliotecas e Ferramentas

- Kit data scientist on Python:
 - Numpy, Scipy, Matplotlib, Pandas, scikit-learn, Pytorch/Tensorflow, Tensorboard, scikit-image e OpenCV.
- Ferramentas:
 - Weka – Biblioteca mais famosa de MD
 - R – Software estatístico
 - OpenCV – Biblioteca para Visão Computacional (C/C++ e Python)
 - Matlab – Software de cálculo numérico (PI)
 - ...

Emprego

USA: 128 mil

USA: 128 mil

Search data science United States

Jobs Date Posted Experience Level Company Job Type Remote Easy Apply All filters

Data science in United States 128,315 results Job Alert Off

Principal Data Science Lead
Microsoft · Redmond, WA
3 connections work here
Promoted · 1 applicant

AI/ML - Machine Learning Engineer, Siri Understanding
Apple
Cambridge, MA
Actively recruiting
Promoted

Senior Researcher, Data Science
Xerox
Cary, NC
Promoted

Principal Data Science Lead
Microsoft · Redmond, WA
Posted 5 days ago · 15 views
Apply Save
See recent hiring trends for this company
Try Premium for free

Job	Company
1 applicant	10001+ employees
Full-time	Computer Software

As cyber-attacks have become more sophisticated, Microsoft Threat Intelligence Center (TIC) enterprises detect, investigate, and respond to advanced attacks and threats.

Canada: 8.058

Canada: 8.058

Search data science Canada

Jobs Date Posted Experience Level Company Job Type Remote Easy Apply All filters

Data science in Canada 8,058 results Job Alert Off

Senior Data Scientist
Microsoft · Vancouver, BC
3 connections work here
Promoted · 0 applicants

Research Scientist, AI
Facebook
Montreal, QC
5 connections work here
Promoted

Data Science Consultant
Avanade
Vancouver, BC
1 company alum works here
Promoted

Senior Data Scientist
Microsoft · Vancouver, BC
Posted 5 days ago · 32 views
Apply Save
See recent hiring trends for this company
Try Premium for free

Job	Company	Connections
0 applicants	10001+ employees	3 connections
Full-time	Computer Software	1 company alum

Microsoft Teams is the hub for teamwork in Office 365 that integrates all the people, content, and tools a team needs to be more engaged and effective. It is core to Microsoft's modern work, modern life & modern education value prop. Nothing can stop a Team!

Australia: 1.747

Australia: 1.747

Search data science Australia

Jobs Date Posted Experience Level Company Job Type Remote Easy Apply All filters

Data science in Australia 1,747 results Job Alert Off

NHMR Grant-Funded Researcher A - School of Computer Science
University of Adelaide
Adelaide, South Australia, Australia
13 connections work here
Promoted · 2 applicants

Senior AI Engineer
Annalise.ai
Sydney, New South Wales, Australia · Remote
Actively recruiting
Promoted

Data Science Lead
IBM
Ballarat, Victoria, Australia
9 connections work here
5 days ago · 0 applicants

NHMR Grant-Funded Researcher A - School of Computer Science
University of Adelaide · Adelaide, South Australia, Australia
Posted 3 days ago · 45 views
Apply Save
See recent hiring trends for this company
Try Premium for free

Job	Company
2 applicants	5,001-10,000 employees
Associate	Higher Education

Germany: 8.340

Germany: 8.340

Search data science Germany

Jobs Date Posted Experience Level Company Job Type Remote Easy Apply All filters

Data science in Germany 8,340 results Job Alert Off

Data Science Lead (d/f/m)
Cognizant
Frankfurt am Main, Hesse, Germany
1 company alum works here
Promoted

Data Science Manager
European Recruitment
Munich, Bavaria, Germany
Actively recruiting
Promoted · Easy Apply

Senior Research Scientist - Data Science team
Amazon
Berlin, Berlin, Germany
7 connections work here
Promoted

Data Science Lead (d/f/m)
Cognizant · Frankfurt am Main, Hesse, Germany
Posted 3 weeks ago · 313 views
Apply Save
See how you compare to 29 applicants
Try Premium Free for 1 Month


Job	Company	Connections
29 applicants	10001+ employees	1 company alum
Entry level	Information Technology...	12 alumni








What makes Cognizant a unique place to work? The combination of rapid growth and an international and innovative environment! This is creating many opportunities for people like YOU — people with an entrepreneurial spirit who want to make a difference

(Level A) \$89,610 per annum plus an employer contribution of 9.5% superannuation

Emprego

Brazil: 1.915





Try Premium Free for 1 Month

Jobs

Date Posted

Experience Level

Company

Job Type

Remote


Easy Apply

All filters

Data science in Brazil

1,915 results


Job Alert Off




Data Science Manager

QuintoAndar

São Paulo, Brazil · Remote

 4 connections work here


Promoted · 3 applicants · Easy Apply




Data Science Lead

Creditas

São Paulo, São Paulo, Brazil

 4 company alumni work here


Promoted




Lead Data Science

PicPay

São Paulo, São Paulo, Brazil · Remote

 1 company alum works here


1 week ago · 12 applicants




Gerente de Data Science

ePharma - Inovação, Integração e Cuidado em Saúde

Barueri, São Paulo, Brazil · Remote

 1 company alum works here


Promoted · Easy Apply




Gerente De Data Science

Neon

São Paulo, São Paulo, Brazil

 1 company alum works here


Promoted · Easy Apply




Head of Data Science

Fanatee

São Paulo, Brazil

 1 alum works here



Data Science Manager

QuintoAndar · São Paulo, Brazil · Remote



Posted 4 days ago · 89 views

Easy Apply

Save

See how you compare to 3 applicants

Try Premium Free for 1 Month

Job	Company	Connections
<ul style="list-style-type: none">3 applicantsMid-Senior level	<ul style="list-style-type: none">1,001-5,000 employeesInternet	<div> 4 connections</div> <div> 4 company alumni</div>

Every Data Scientist at QuintoAndar

In our Data Science team, you'll have the chance to work as a part of our product squads delivering features that enrich our product and a first class experience that we are known for. You'll work alongside brilliant minds, with different skills: Product Managers, Designers, UX Writers, Data Scientists, Machine Learning Engineers, Software Engineers, etc. Be prepared for the job of your life - hard challenges, high expectations, intelligent life at work, meaningful conversations, outstanding productivity.

Some of the specific problems we work on the Data Science team include intelligent search and recommendation systems; customer experience enhancement using ML such as in automatic routing of tickets; pricing, liquidity and credit models; building a platform that enables us to develop and deploy all those solutions.


Data Science Manager responsibilities at QuintoAndar

On top of being a great Data Scientist, we expect you to

- Lead a cross-functional team of 3-6 Data Scientists and Software Engineers to build data products to solve business challenges;
- Be a reference and build a great team by both recruiting great talent and helping them grow;
- Discuss business requirements with the Product Manager on a deep level and

Emprego

Brazil: 1.915



Jobs

Date Posted

Experience Level

Company

Job Type

Remote


Easy Apply


All filters


Data science in Brazil


1,915 results


Job Alert Off


**Data Science Manager**
QuintoAndar
São Paulo, Brazil · Remote
4 connections work here
Promoted · 3 applicants · Easy Apply


**Data Science Lead**
Credits
São Paulo, São Paulo, Brazil
4 company alumni work here
Promoted

**Lead Data Science**
PicPay
São Paulo, São Paulo, Brazil · Remote
1 company alum works here
1 week ago · 12 applicants

**Gerente de Data Science**
ePharma - Inovação, Integração e Cidadania em Saúde
Barueri, São Paulo, Brazil · Remote
1 company alum works here
Promoted · Easy Apply

**Gerente De Data Science**
Neon
São Paulo, São Paulo, Brazil
1 company alum works here
Promoted · Easy Apply

**Head of Data Science**
Fanatee
São Paulo, Brazil
1 alum works here

**Data Science Manager**
QuintoAndar · São Paulo, Brazil · Remote
Posted 4 days ago · 89 views
Easy Apply · Save
Schedule to compare to 3 applicants
Try Premium Free for 1 Month

Job	Company	Connections
10 applicants Mid-Senior level	• 1,001-5,000 employees • Internet	4 connections 4 company alumni

Every Data Scientist at QuintoAndar

In our Data Science team, you'll have the chance to work as a part of our product squads delivering features that enrich our product and a first class experience that we are known for. You'll work alongside brilliant minds, with different skills: Product Managers, Designers, UX Writers, Data Scientists, Machine Learning Engineers, Software Engineers, etc. Be prepared for the job of your life - hard challenges, high expectations, intelligent life at work, meaningful conversations, outstanding productivity.

Some of the specific problems we work on the Data Science team include intelligent search and recommendation systems; customer experience enhancement using ML such as in automatic routing of tickets; pricing, liquidity and credit models; building a platform that enables us to develop and deploy all those solutions.

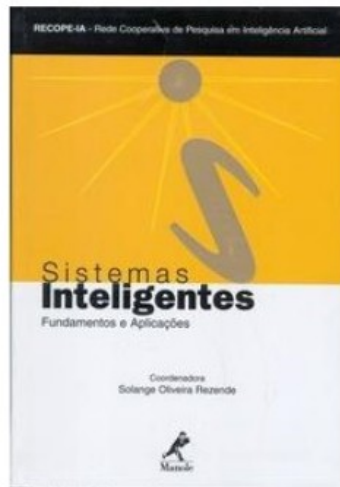
Data Science Manager responsibilities at QuintoAndar

On top of being a great Data Scientist, we expect you to

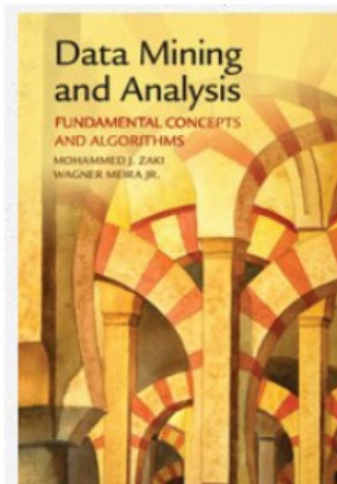
- Lead a cross-functional team of 3-6 Data Scientists and Software Engineers to build data products to solve business challenges;
- Be a reference and build a great team by both recruiting great talent and helping them grow;
- Discuss business requirements with the Product Manager on a deep level and

Referências

- 1) **Fayyad et al.** From data mining to knowledge discovery: an overview. In Advances in Knowledge Discovery & Data Mining, pp. 1–34, 1996.
- 2) **Rezende et al.** Sistemas inteligentes: fundamentos e aplicações, 2003.
- 3) **Zaki and Meira.** Data Mining and Analysis: Fundamental Concepts and Algorithms, May 2014.



[2]



[3]

Livros

According to Data Science Books You Must Read in 2020:

- Introduction to Statistical Learning
- Deep Learning with Python
- Deep Learning
- ...
- ...
- ...

<https://towardsdatascience.com/data-science-books-you-must-read-in-2020-1f30daace1cb>