PÓS-GRADUAÇÃO EM CIÊNCIA DE DADOS E BIG DATA

IEC PUC MINAS

Disciplina:

MODELAGEM E PREPARAÇÃO DE DADOS PARA APRENDIZADO DE MÁQUINA

Professor: Rafael Barroso

PUC Minas

PUC Minas

APRESENTAÇÃO

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS

2014:
Engenheiro de Minas

USP
Universidade de São Paulo

2018:
MSc. Engenharia Mineral

ITA

2018:
Especialização em Ciência de Dados

FIA BUSINESS SCHOOL

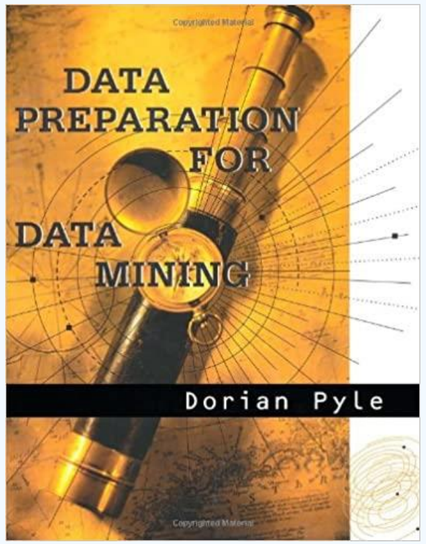
2018:
Especialização em Finanças Corporativas e Investment Banking

Itaú

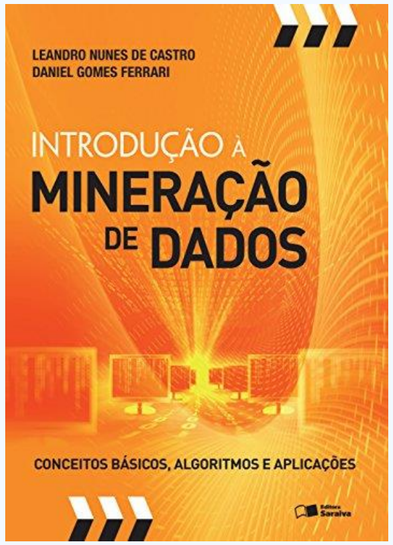
PUC Minas

BIBLIOGRAFIA E REFERÊNCIAS

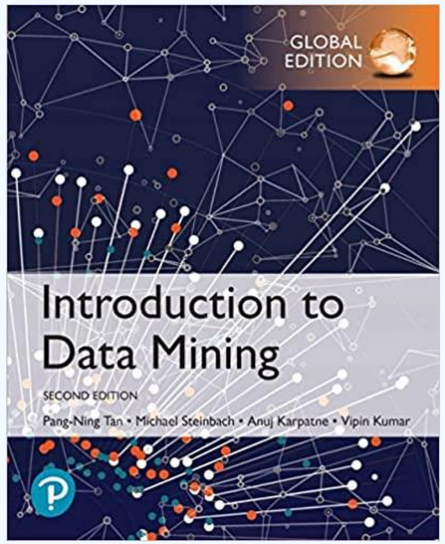
CONCEITOS BÁSICOS



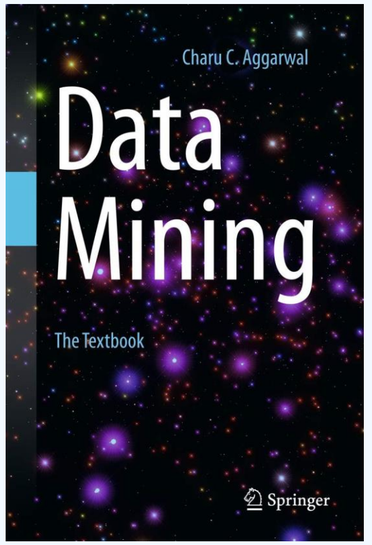
Data Preparation for Data Mining
PYLE, 1999



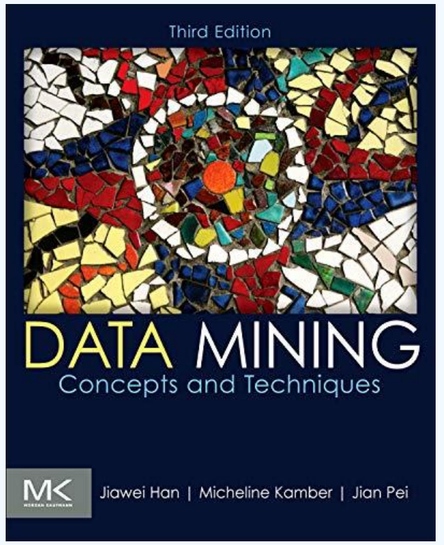
Introdução à Mineração de Dados
CASTRO e FERRARI, 2017



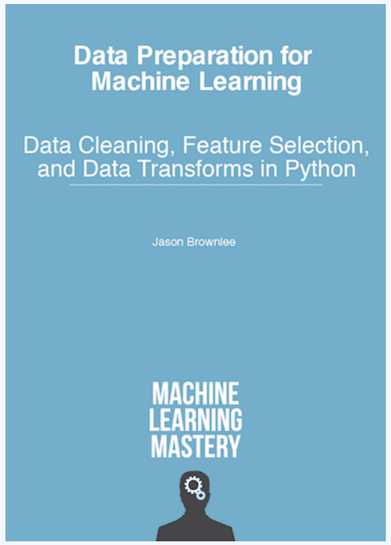
Introduction to Data Mining
TAN, STEINBACH, KARPATNE e KUMAR, 2019



Data Mining – The Textbook
AGGARWAL, 2015



Data Mining – Concepts and Techniques
HAN, KAMBER e PEI, 2011

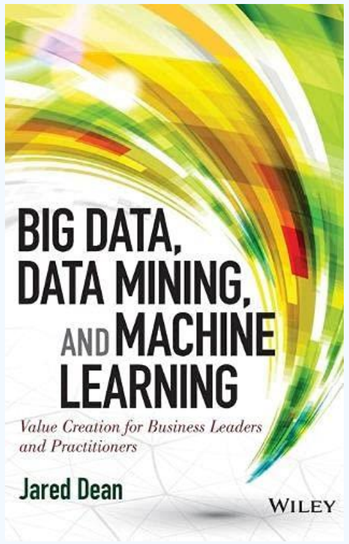


Data Preparation for Machine Learning
BROWNLEE, 2020

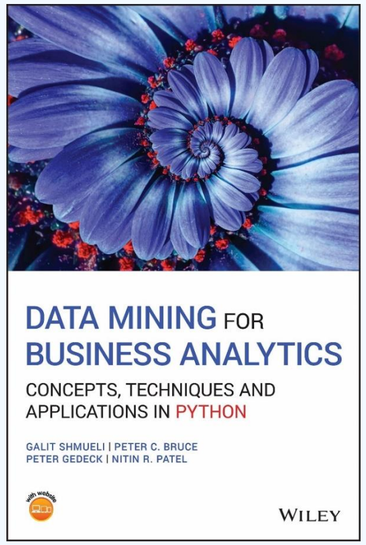
PUC Minas

BIBLIOGRAFIA E REFERÊNCIAS

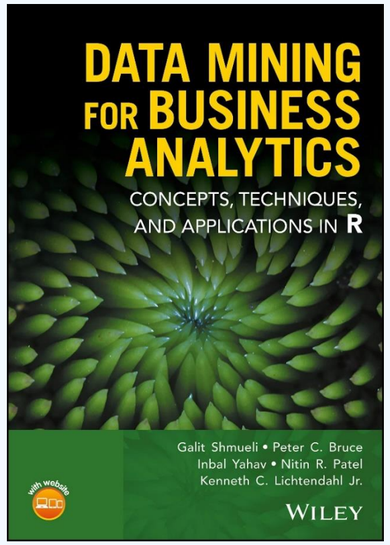
DATA MINING E NEGÓCIOS



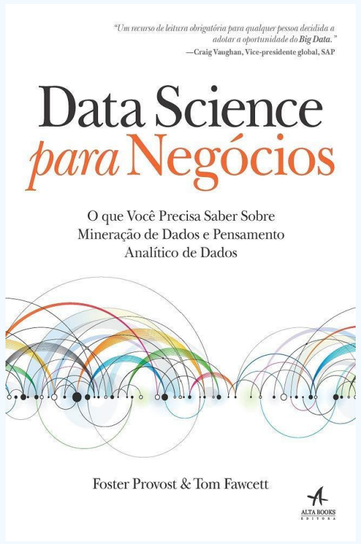
Big Data, Data Mining and Machine Learning
DEAN, 2014



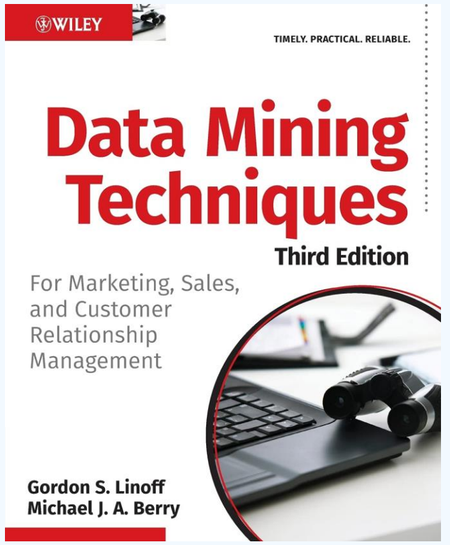
Data Mining for Business Analytics – Concepts, Techniques and Applications in Python
SHMUELI, BRUCE, GEDECK e PATEL, 2020



Data Mining for Business Analytics – Concepts, Techniques and Applications in R
SHMUELI, BRUCE, YAHAV, PATEL e LICHTENDAHL JR., 2017



Data Science para Negócios
PROVOST e FAWCETT, 2016



Data Mining Techniques for Marketing, Sales, and Customer Relationship Management
LINOFF e BERRY, 2011

PUC Minas

PARTE 1

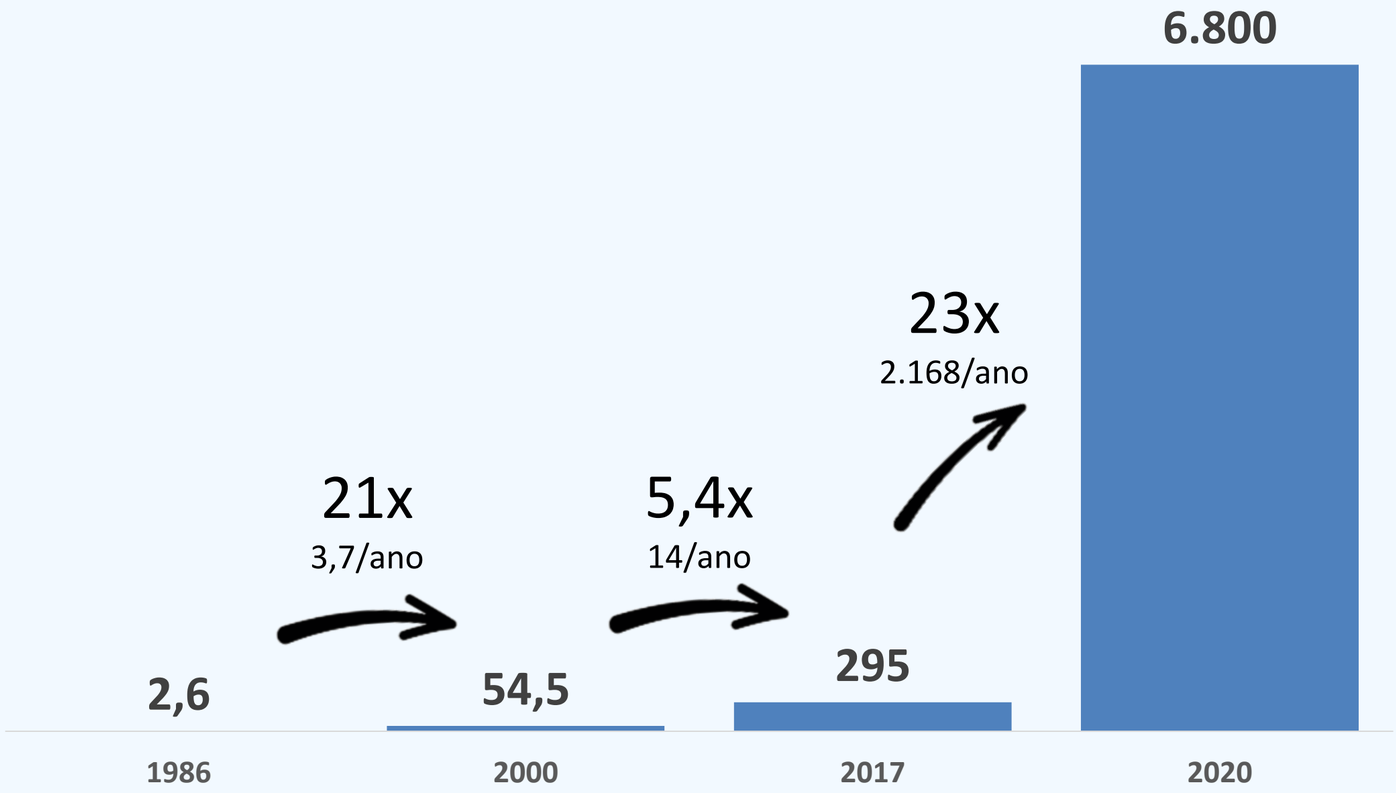
DADO, INFORMAÇÃO E CONHECIMENTO

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

SERÁ QUE TODO DADO É ÚTIL OU UTILIZÁVEL?

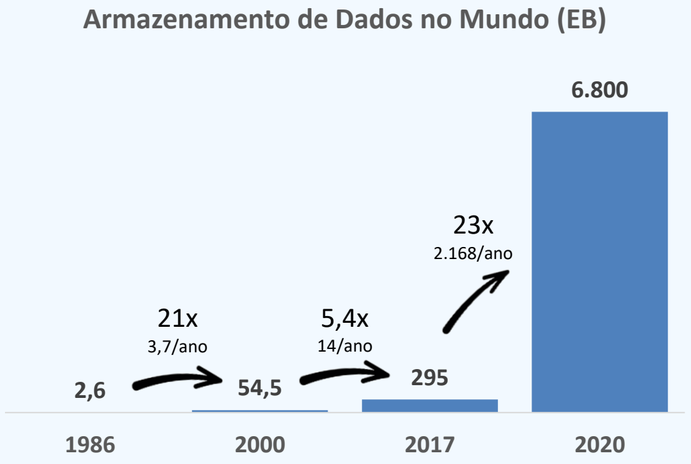
1 EB = 10^{18}B

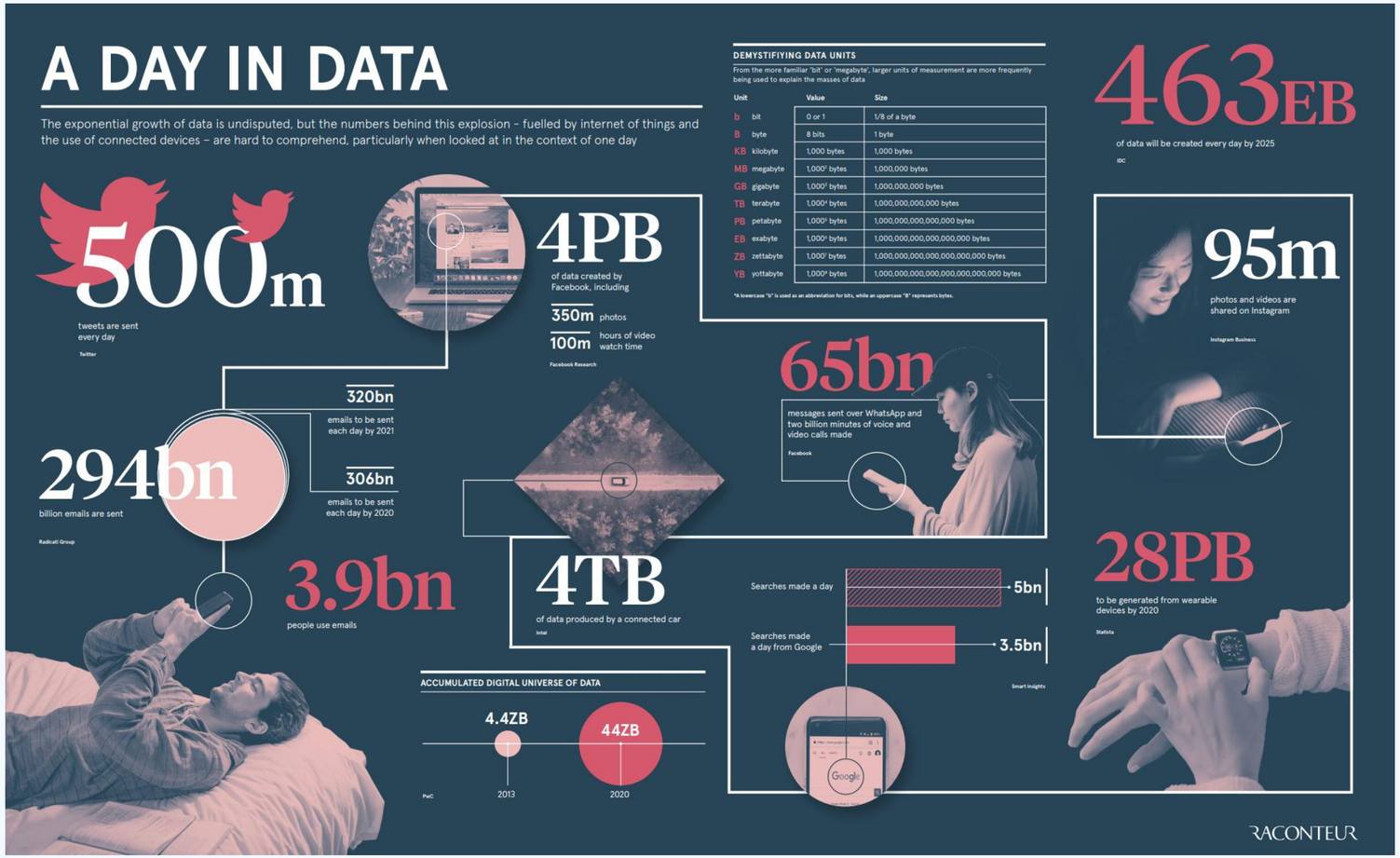


PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

SERÁ QUE TODO DADO É ÚTIL OU UTILIZÁVEL?

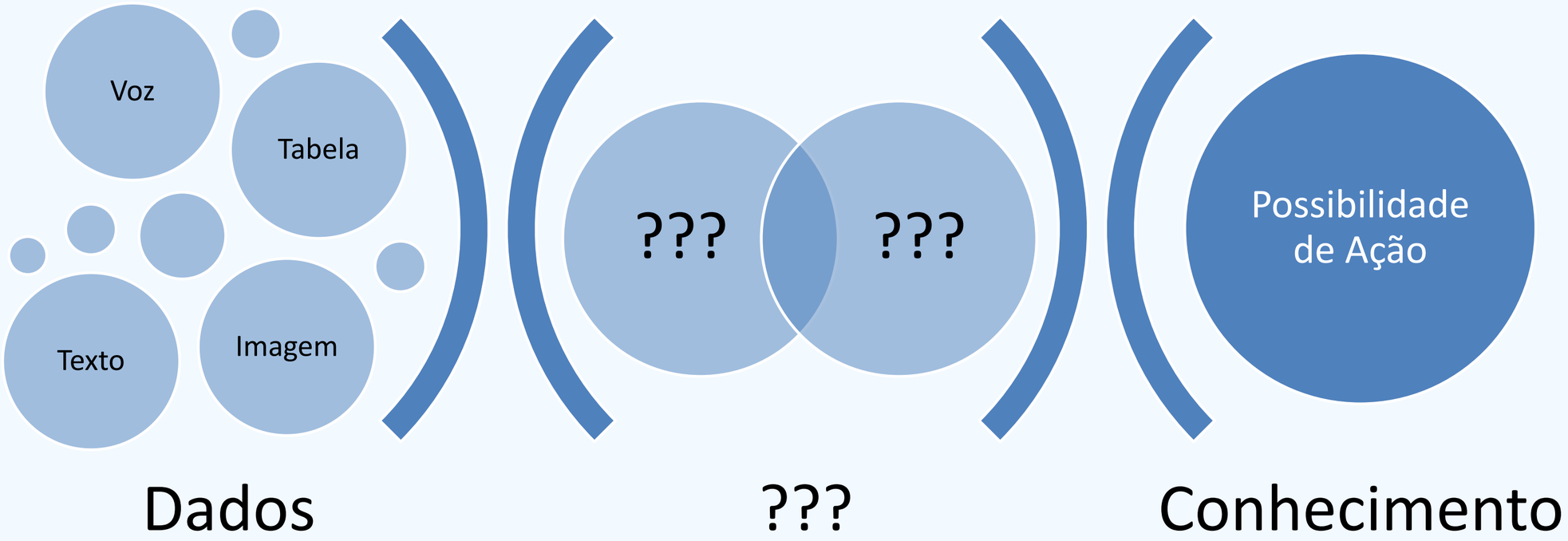




Fonte: https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-world/

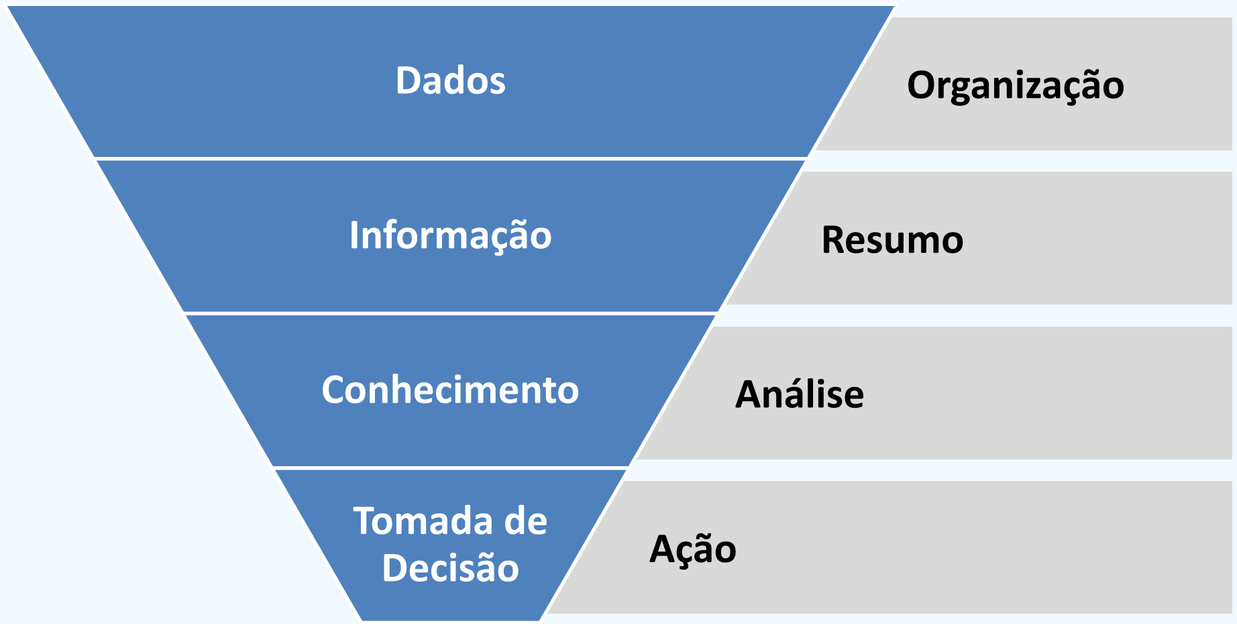
PUC Minas

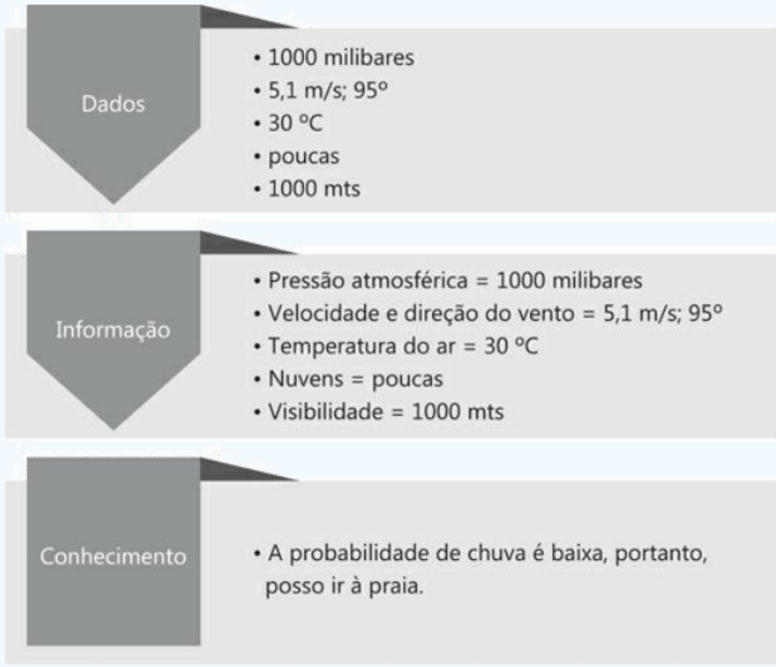
DADO, INFORMAÇÃO E CONHECIMENTO



PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO





Fonte: Introdução à Mineração de Dados, CASTRO e FERRARI, 2017

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

PUC Minas

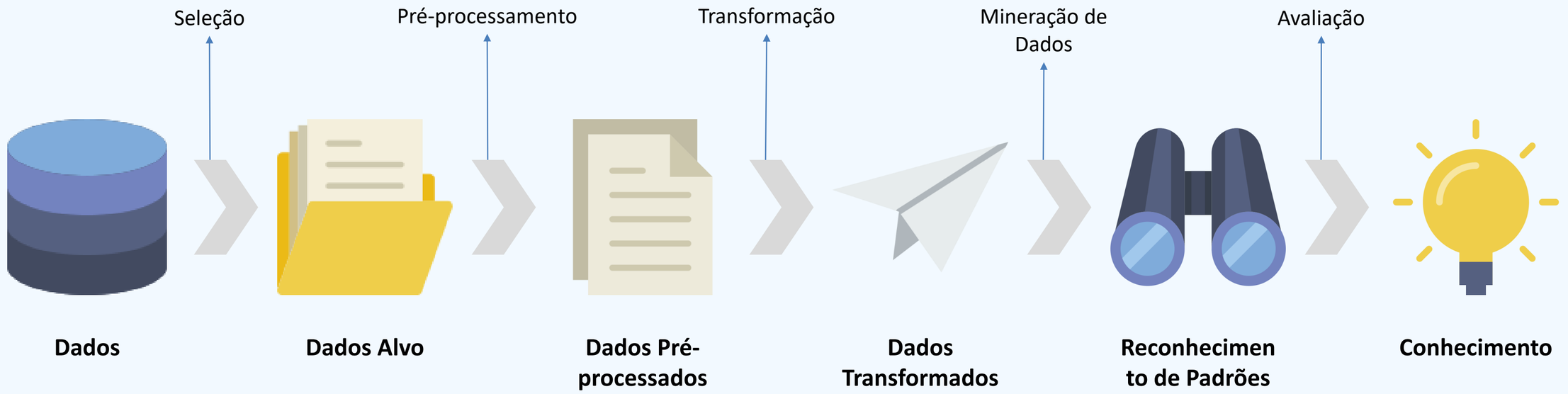
DADO, INFORMAÇÃO E CONHECIMENTO

- A mineração de dados faz parte de algo maior chamado *descoberta de conhecimento em bases de dados* (do inglês: *knowledge discovery in databases*, KDD) ;

- KDD se divide em:
  - Seleção e integração das bases de dados,
  - Limpeza da base,
  - Seleção e transformação de dados,
  - Mineração,
  - Avaliação de dados.

PUC Minas

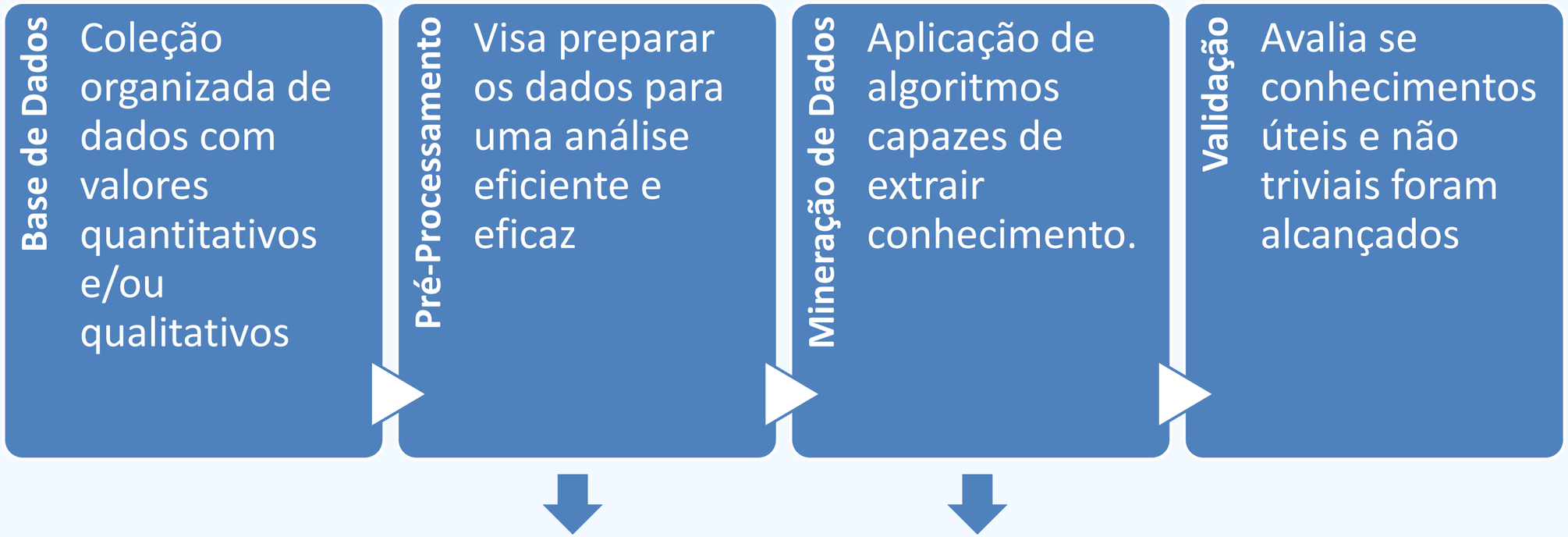
DADO, INFORMAÇÃO E CONHECIMENTO



Esquema geral do processo de descobrimento de conhecimento em bases de dados (KDD)

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

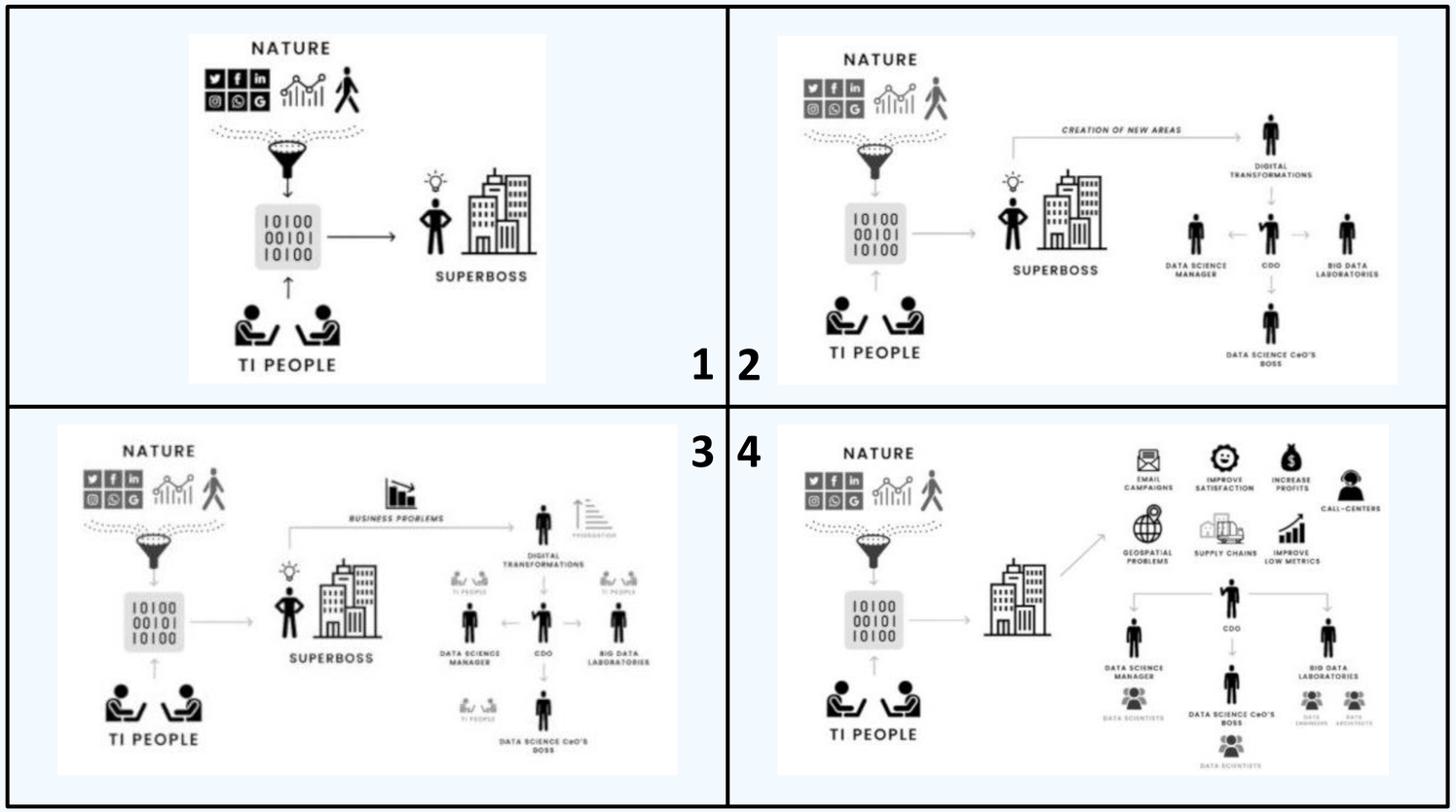


Foco (individual) do ciêntista de dados!

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

Primeiros passos de um projeto de ciência de dados (KDNuggets, 2020):

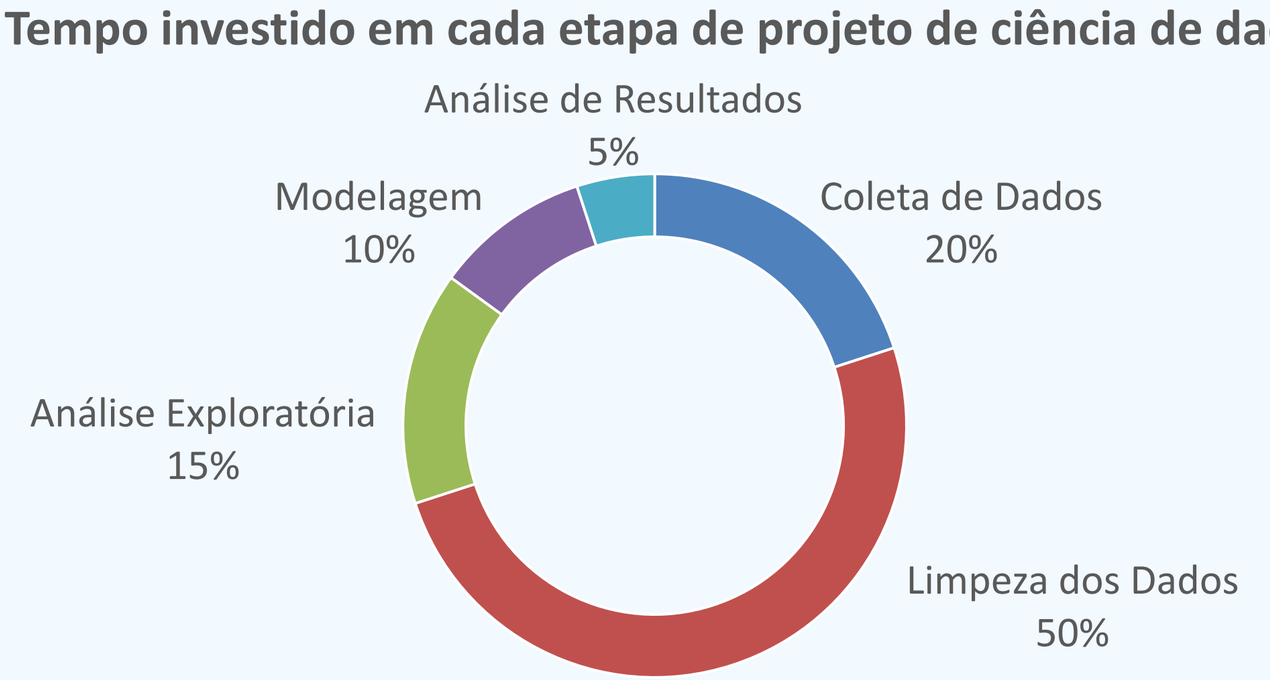


1. “Temos um monte de dados, precisamos fazer algo com eles!”
2. Criação de novas áreas para pensar em como usar os dados + novas contratações;
3. Projetos de ciência de dados são definidos e priorizados;
4. Solução de desafios usando dados , *machine learning* e análise estatística. Casos de uso.

Fonte: https://www.kdnuggets.com/2020/07/first-steps-data-science-project.html

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

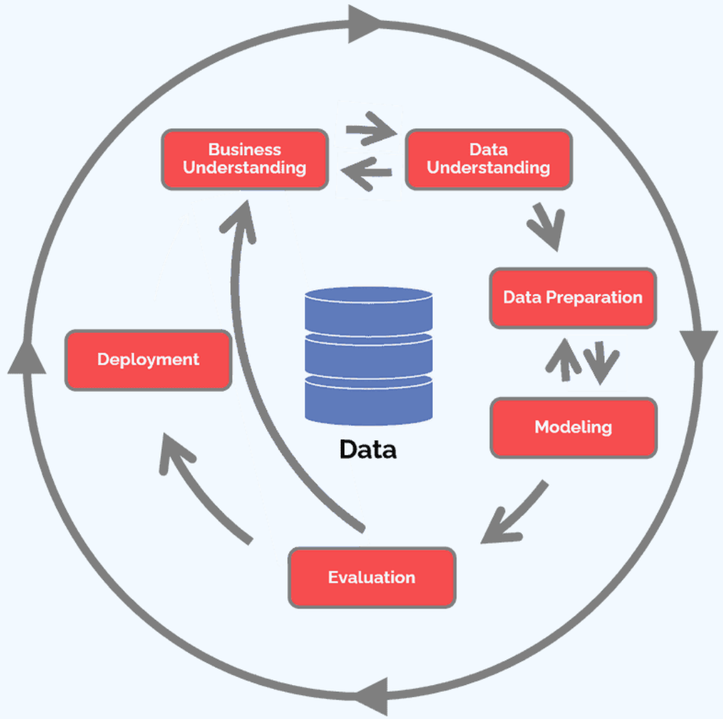


Fonte: https://anyinstructor.com/what-is-the-average-duration-of-a-data-science-project/#:~:text=It%20will%20take%20between%202,and%20needs%20of%20the%20project.

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

CRISP-DM: ***Cross Industry Standard Process for Data Mining***



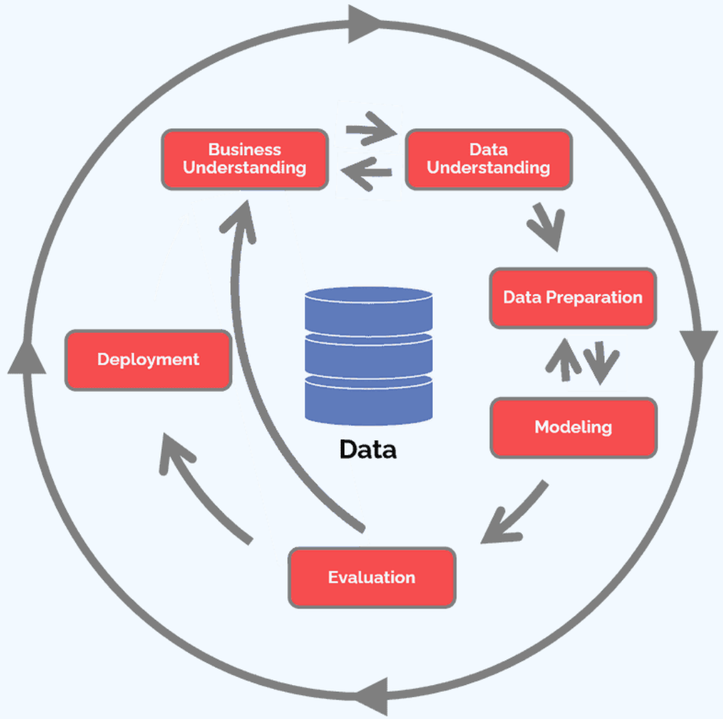
- Descreve naturalmente o ciclo de vida de um projeto de ciência de dados;
- Inicia todo o processo com Entendimento de Negócio;
- Esta primeira etapa é dividida em:
  - Determinação de objetivos do negócio,
  - Avaliar a realidade atual,
  - Determinar os objetivos da mineração de dados,
  - Planejar o projeto;
- Bom conhecimento de negócio é essencial para avaliar a solução.

Fonte: https://www.datascience-pm.com/crisp-dm-2/

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

CRISP-DM: ***Cross Industry Standard Process for Data Mining***



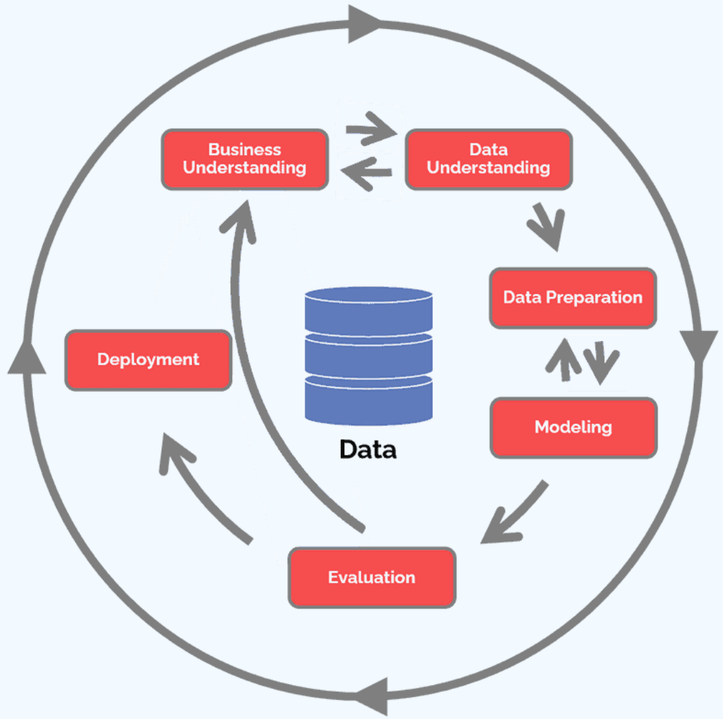
Fonte: https://www.datascience-pm.com/crisp-dm-2/

- A segunda etapa é o Entendimento dos Dados, dividida em:
  - Coleta inicial de dados,
  - Descrição dos dados,
  - Análise exploratória,
  - Verificação de qualidade;
- A terceira etapa é a Preparação dos Dados, que se subdivide em:
  - Seleção,
  - Limpeza,
  - Construção,
  - Integração,
  - Formatação;

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

CRISP-DM: ***Cross Industry Standard Process for Data Mining***



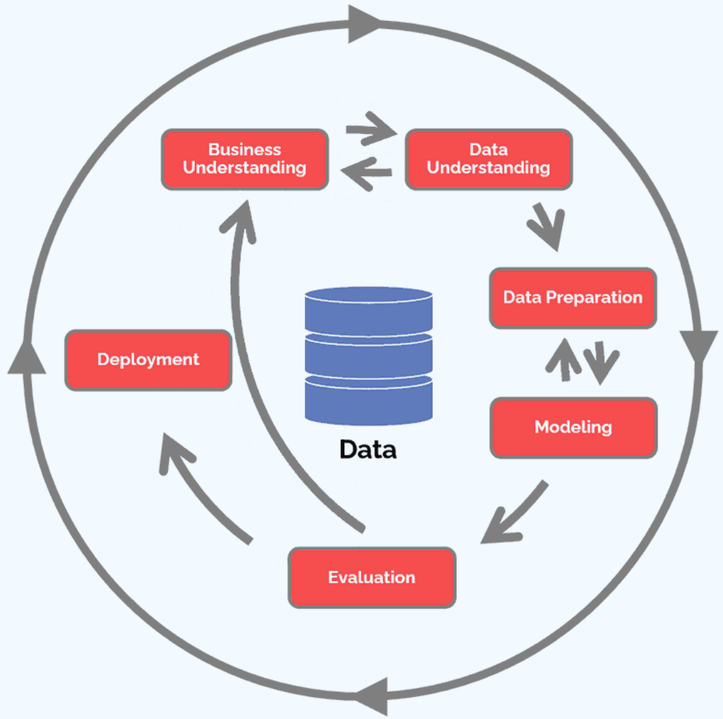
- A Modelagem é a terceira etapa e divide-se em:
  - Seleção de técnicas,
  - Criação de base de teste,
  - Construção do modelo,
  - Avaliação do modelo;
- Os subitens da quarta etapa, a Avaliação, são:
  - Avaliação de resultados baseado nos critérios de negócio,
  - Revisão do processo,
  - Determinação de próximos passos;

Fonte: https://www.datascience-pm.com/crisp-dm-2/

PUC Minas

DADO, INFORMAÇÃO E CONHECIMENTO

CRISP-DM: ***Cross Industry Standard Process for Data Mining***



- O último estágio é a Implementação, dividida em:
  - Planejamento da operacionalização,
  - Planejamento da monitoração e manutenção,
  - Documentação final,
  - Revisão para próximos projetos.

Fonte: https://www.datascience-pm.com/crisp-dm-2/

PUC Minas

PARTE 2

DEFININDO O DOMÍNIO DO PROBLEMA

DEFININDO O DOMÍNIO DO PROBLEMA

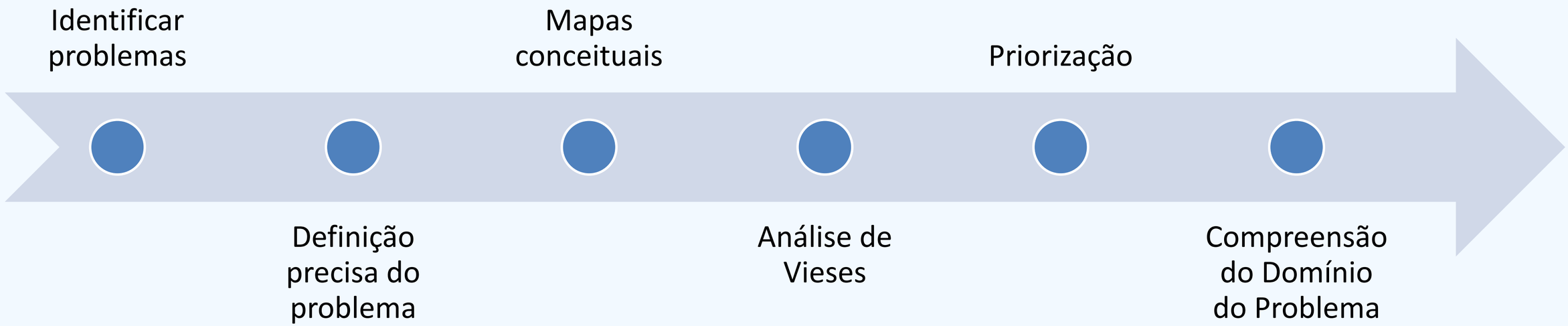
- Antes de sair buscando uma solução mágica (**que não existe!**) nos dados, precisamos saber encontrar e definir um problema para solucionar.
- Parece fácil, mas não é. Por exemplo: por que existe desemprego?
- Falta de oportunidade? Tem gente que não quer trabalhar? Má formação?
- Não importa o “porquê”, cada um demanda uma solução específica, muito bem planejada e executada para dar certo.

DEFININDO O DOMÍNIO DO PROBLEMA

- Trabalhar com dados é igualmente complexo, buscar uma solução simplista geralmente vai induzir ao erro.
- Toda resposta é tão boa quanto a qualidade da pergunta que ela responde.
- Quanto mais específica (bem delimitada) for a pergunta, maior a chance de encontrar uma resposta.
- Mesmo quando não se encontra uma resposta, temos uma resposta.

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA



PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

- Existem **vários** problemas, mas qual é o problema?
- Se temos tanta certeza de qual é o problema, não absorvemos o que os dados mostram (as vezes, nem olhamos direito pra eles!).

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

- Tentar trabalhar com um problema “pré-pronto” pode não ser o que, de fato, gera valor.
- Então como encontrar o problema?

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

- A melhor solução: ser específico!
- Buscar perguntas que podem ter suas respostas descobertas no dados.
- (-) “Quem vai cancelar o contrato?”
- (+) “Entre os clientes com até um ano de contrato e que tenham cancelado o contrato, qual o comportamento de consumo?”

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

- Quanto mais complexo for o domínio do problema, mais importante visualizar o processo.
- Mapear para entender a cadeia de valor (e dados) do assunto no qual vamos atuar.
- A ideia é entender como as coisas se conectam!

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

Fonte: https://www.goconqr.com/pt-BR/examtime/blog/tudo-sobre-mapas-mentais/

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

- O que está na cabeça do “dono do problema”?
- Como garantir que quem vai criar a solução entendeu?
- É essencial que ambos expressem suas suposições/hipóteses sobre o problema e estejam alinhados.

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

- Podemos descobrir e definir vários problemas no decorrer do processo.
- É necessário saber qual priorizar.
- O mais fácil? O que traz mais retorno financeiro? O que impacta mais clientes?

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

Análise Par-a-Par (*Pairwise Analysis*):

Qual melhor filme?
Harry Potter
Star Wars
Rei Leão
Senhor dos Anéis
Jogos Vorazes
Jurassic Park
Vingadores

	HP	SW	RL	SA	JV	JP	V	Resultado
HP	-							
SW		-						
RL			-					
SA				-				
JV					-			
JP						-		
V							-	

Fonte: https://www.1000minds.com/decision-making/pairwise-comparison

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

Análise Par-a-Par (*Pairwise Analysis*):

Qual melhor filme?
Harry Potter
Star Wars
Rei Leão
Senhor dos Anéis
Jogos Vorazes
Jurassic Park
Vingadores

	HP	SW	RL	SA	JV	JP	V	Resultado
HP	-	1	0	1	0	0	0	2
SW	0	-	0	1	0	1	1	3
RL	1	1	-	1	0	0	1	4
SA	0	0	0	-	0	1	1	2
JV	1	1	1	1	-	1	0	5
JP	1	0	1	0	0	-	1	3
V	1	0	0	0	1	1	-	3

Fonte: https://www.1000minds.com/decision-making/pairwise-comparison

PUC Minas

DEFININDO O DOMÍNIO DO PROBLEMA

- Não temos uma fórmula correta que sempre deve ser aplicada.
- Cada caso é um caso.
- O importante é entender, de forma bem definida, qual o desafio e qual seu impacto.

DEFININDO O DOMÍNIO DO PROBLEMA

MAS E A SOLUÇÃO?

- Aqui a grande questão é: o que é uma solução?
- Um problema específico, demanda uma solução específica.
- Não adianta propor uma solução que não seja implementável.
- Gestão de expectativas. Não existe “bala de prata”: a regra é que cada problema tem sua própria solução.

PUC Minas

PARTE 3

CARACTERIZAÇÃO DOS DADOS

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- Dados sempre estão divididos entre detalhados ou agregados.
- Mesmo dados detalhados podem ter origem num agregado de dados menores.
- O nível de granularidade no *dataset* determina o nível de detalhe que ele pode gerar como saída.

- Ex: Saldo Mensal < Saldo Diário < Operações Diárias.

Domínio

Padrões

Integridade

Simultaneidade

Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- A mesma informação pode ser representada de maneiras diferentes.
- Tudo depende do contexto onde a informação é gerada, capturada ou armazenada.

- Ex: temperature em Celsius e Farenheint, mês por escrito ou numeral.

Domínio

Padrões

Integridade

Simultaneidade

Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- Ocorre, geralmente, quando uma informação diferente daquela planejada aparece/é inserida, mas, no contexto, pode até fazer “sentido”.
- Pode ocorrer ainda por transposição accidental na captura/leitura dos dados.
- Resistência humana no preenchimento dos dados.

- Ex: usuário final inclui um valor diferente do esperado, valor “aleatório” que é incluído porque sabe-se que funciona.

Domínio

Padrões

Integridade

Simultaneidade

Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- O que está armazenado naquele dado?
- O objeto que é medido/observado precisa estar bem definido para que o dado seja interpretado corretamente.

- Ex: informação sobre cliente ativo ou inativo.

Domínio

Padrões

Integridade

Simultaneidade

Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

Domínio

Padrões

Integridade

Simultaneidade

Redundância

- Dados oriundos de diferentes fluxos precisam ter seu atributo de relação bem definido.
- Além disso, deve-se garantir que esse atributo esteja definido da mesma maneira nas origens para que junção dos dados seja viável.

- Ex: CPF, CNPJ, número de contrato.

PUC Minas

CARACTERIZAÇÃO DOS DADOS

- Granularidade
- Consistência
- Poluição
- Objetos
- Relações

- Todo atributo tem seu domínio de valores possível pré-definido.
- O domínio pode ser, inclusive, influenciado por valores presentes em outros atributos.

- Ex: idade sempre maior ou igual a zero, profissão e escolaridade

- **Domínio**
- Padrões
- Integridade
- Simultaneidade
- Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- No processo de captura de dados, pode haver a definição de valores padrão para atributos.
- Este valor pode, ainda, ser condicionado por valores em outras variáveis.

- Ex: valor negativo para indicar valor nulo, nome do conjuge vazio quando estado civil é solteiro.

Domínio

Padrões

Integridade

Simultaneidade

Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- Relaciona-se com os limites aceitáveis para dada variável.
- Ao se pensar nesses “limites”, estamos assumindo a existência de valores que podem estar fora deles: os *outliers*.

- Ex: Renda diferente de zero, idade elevada.

Domínio

Padrões

Integridade

Simultaneidade

Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- Relaciona-se à ocorrência de diferentes informações no mesmo período de tempo.
- Pode, ainda, estar relacionado à validade do dado.

- Ex: exame de sangue de 1 ano atrás para prever nível de nutrição atual.

Domínio

Padrões

Integridade

Simultaneidade

Redundância

PUC Minas

CARACTERIZAÇÃO DOS DADOS

Granularidade

Consistência

Poluição

Objetos

Relações

- Ocorre quando informações “iguais” preenchem diferentes atributos.
- Quando essa relação não é óbvia, pode-se ainda identificá-la ao se estudar a colinearidade das variáveis.

- Ex: idade e data de nascimento

Domínio

Padrões

Integridade

Simultaneidade

Redundância

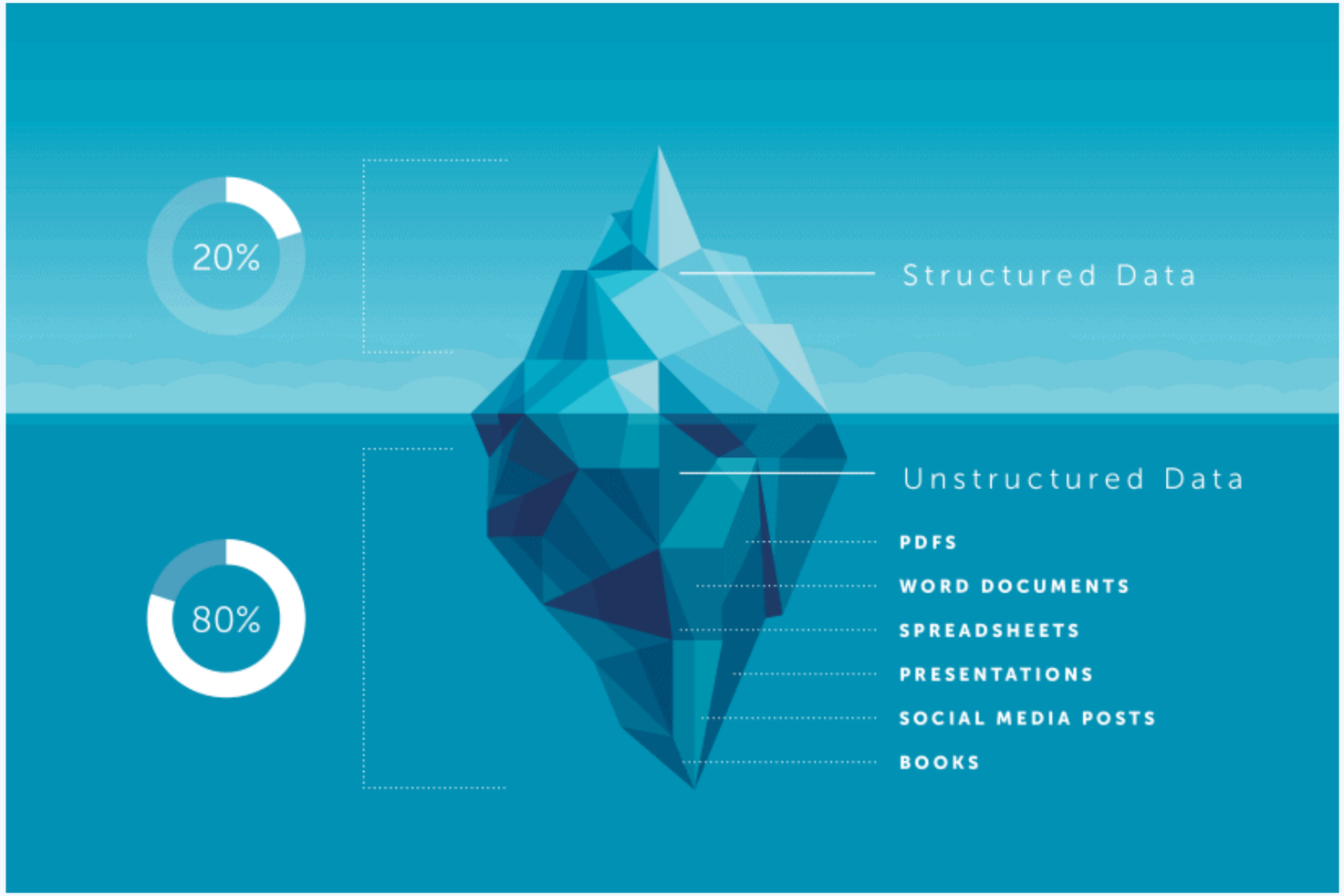
PUC Minas

PARTE 4

TIPOS DE DADOS

PUC Minas

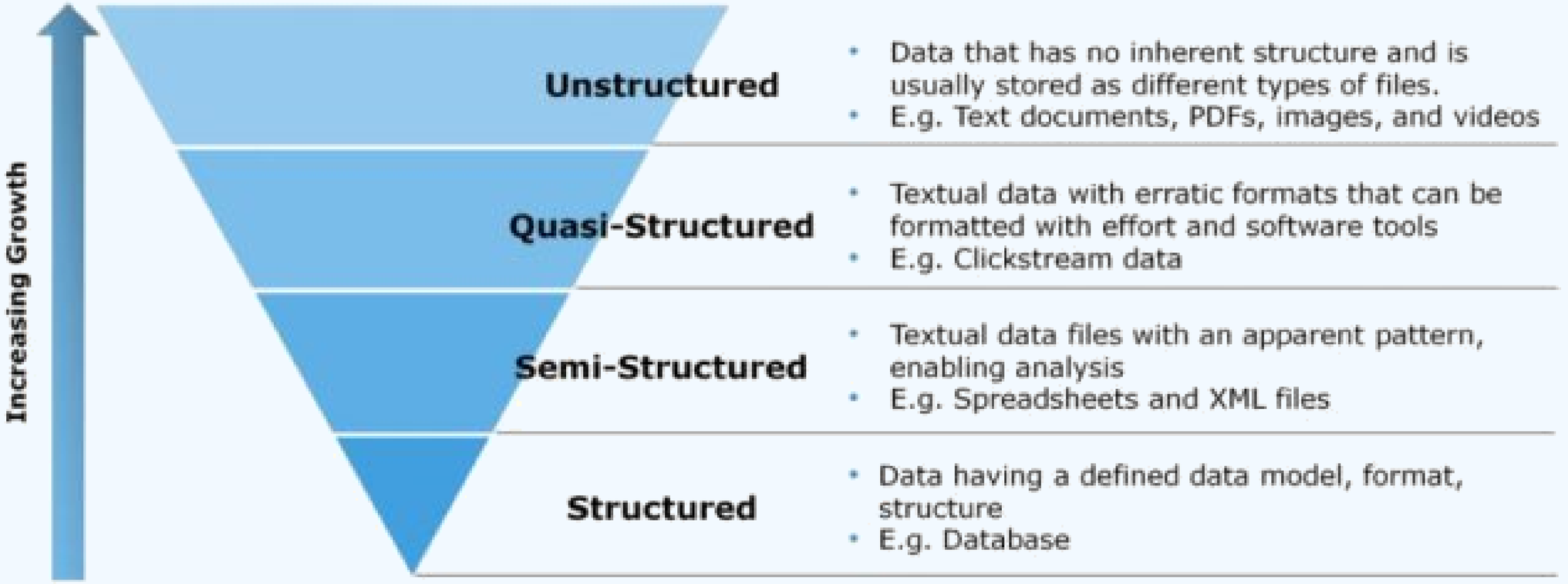
TIPOS DE DADOS



Fonte: https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/

PUC Minas

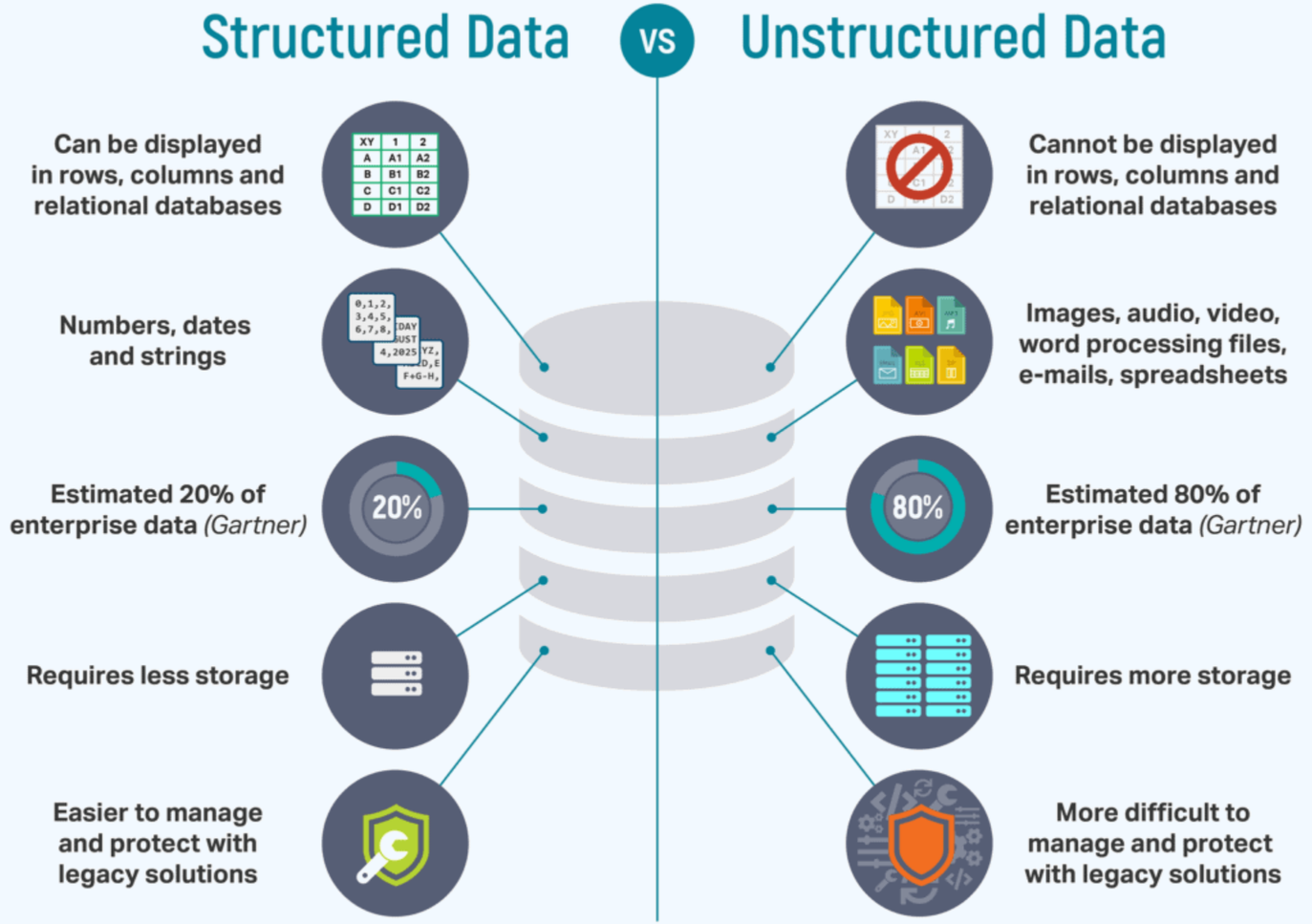
TIPOS DE DADOS



Fonte: https://www.mycloudwiki.com/san/data-and-information-basics/

PUC Minas

TIPOS DE DADOS



Fonte: https://k21academy.com/microsoft-azure/dp-900/structured-data-vs-unstructured-data-vs-semi-structured-data/