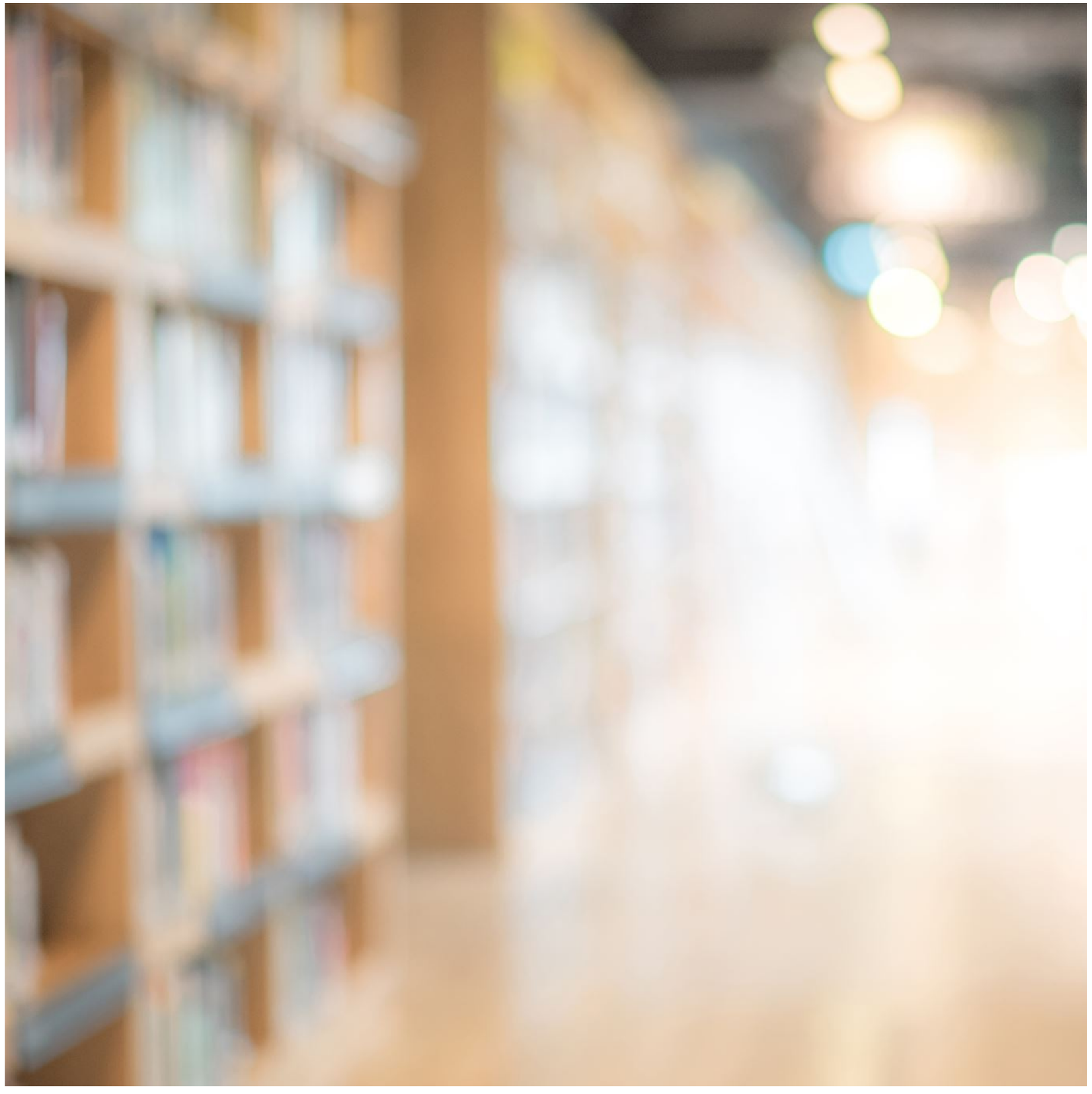


Princípios e Aplicações de Mineração de Dados

Prof. Dra. Karina S. Machado

PPGCOMP

C3

FURG

Ementa

KDD/Ciência de Dados

Pre processamento

Técnicas de Mineração de dados

Tipos de aprendizado/Principais algoritmos

Pós processamento

Interpretação de modelos

Principais Referencias

- TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839 p.

- FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

- HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

- Artigos recentes sobre a área

IA – Machine learning – Data Science – KDD - Data mining – Big Data

- Atualmente não há um consenso entre os pesquisadores da definição exata de cada um destes termos.

- Não tem uma resposta certa ou errada a respeito dessa "diferença"

- Mas vamos a algumas definições e alguns infográficos para discutir e refletir sobre o tema

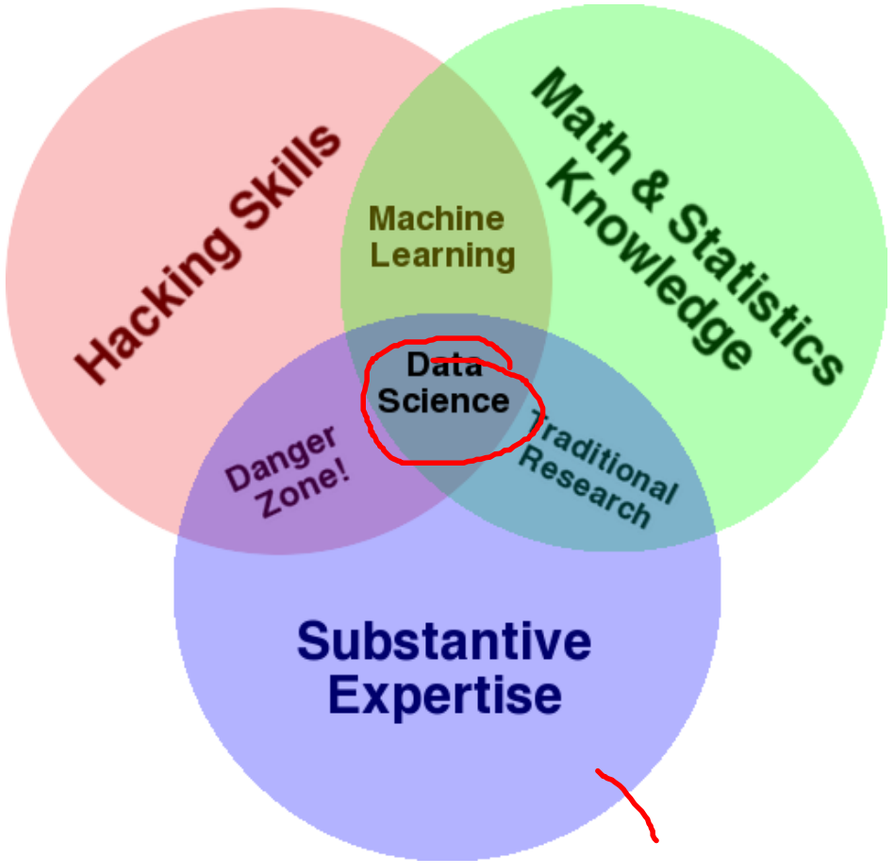
KDD

Data Science

Big Data

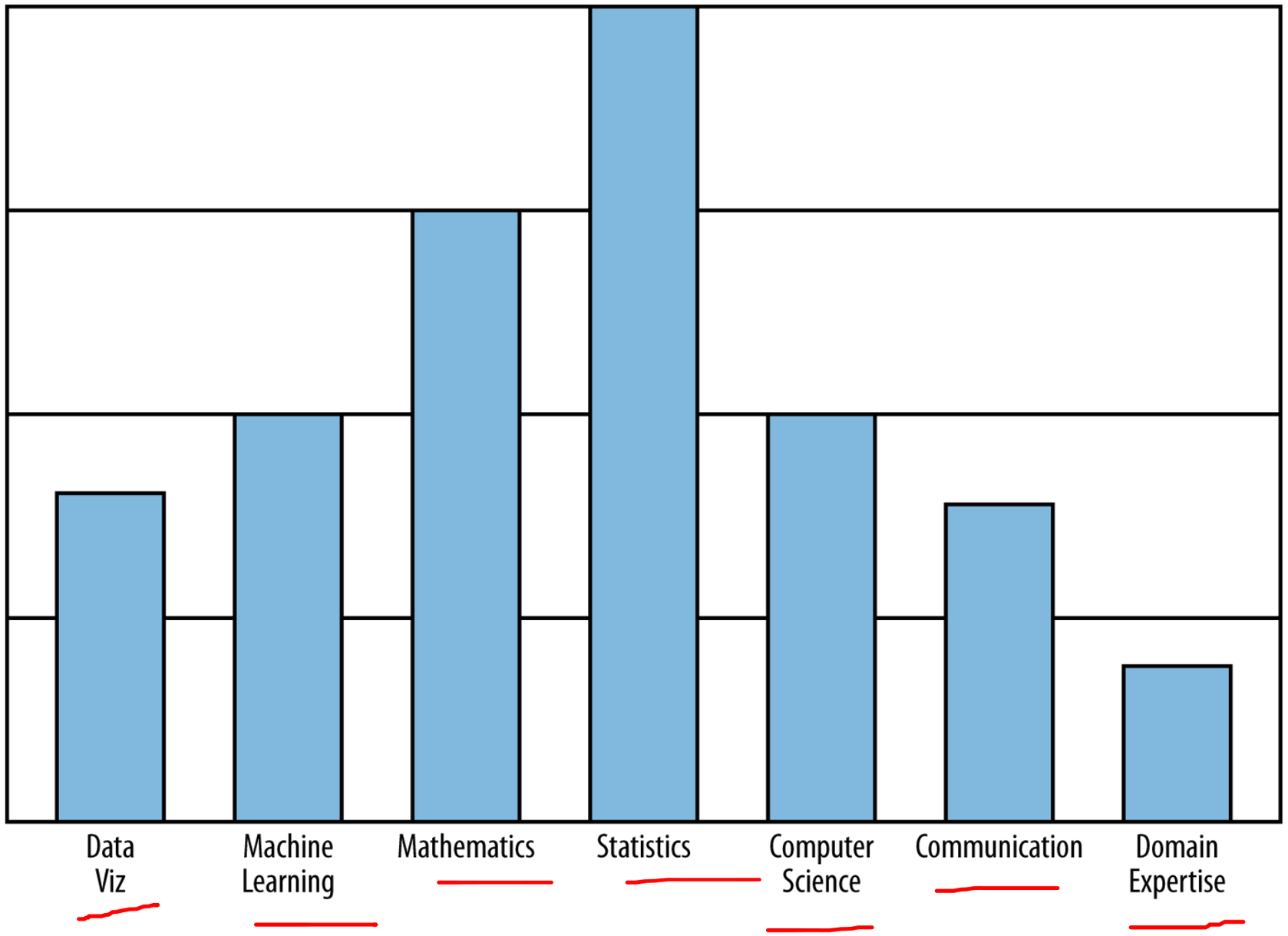
Data Science

- So, what is data science? Is it new, or is it just statistics or analytics rebranded? Is it real, or is it pure hype? And if it's new and if it's real, what does that mean?



O'Neil, Cathy, and Rachel Schutt. *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.", 2013.

https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html



O'Neil, Cathy, and Rachel Schutt. *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.", 2013.

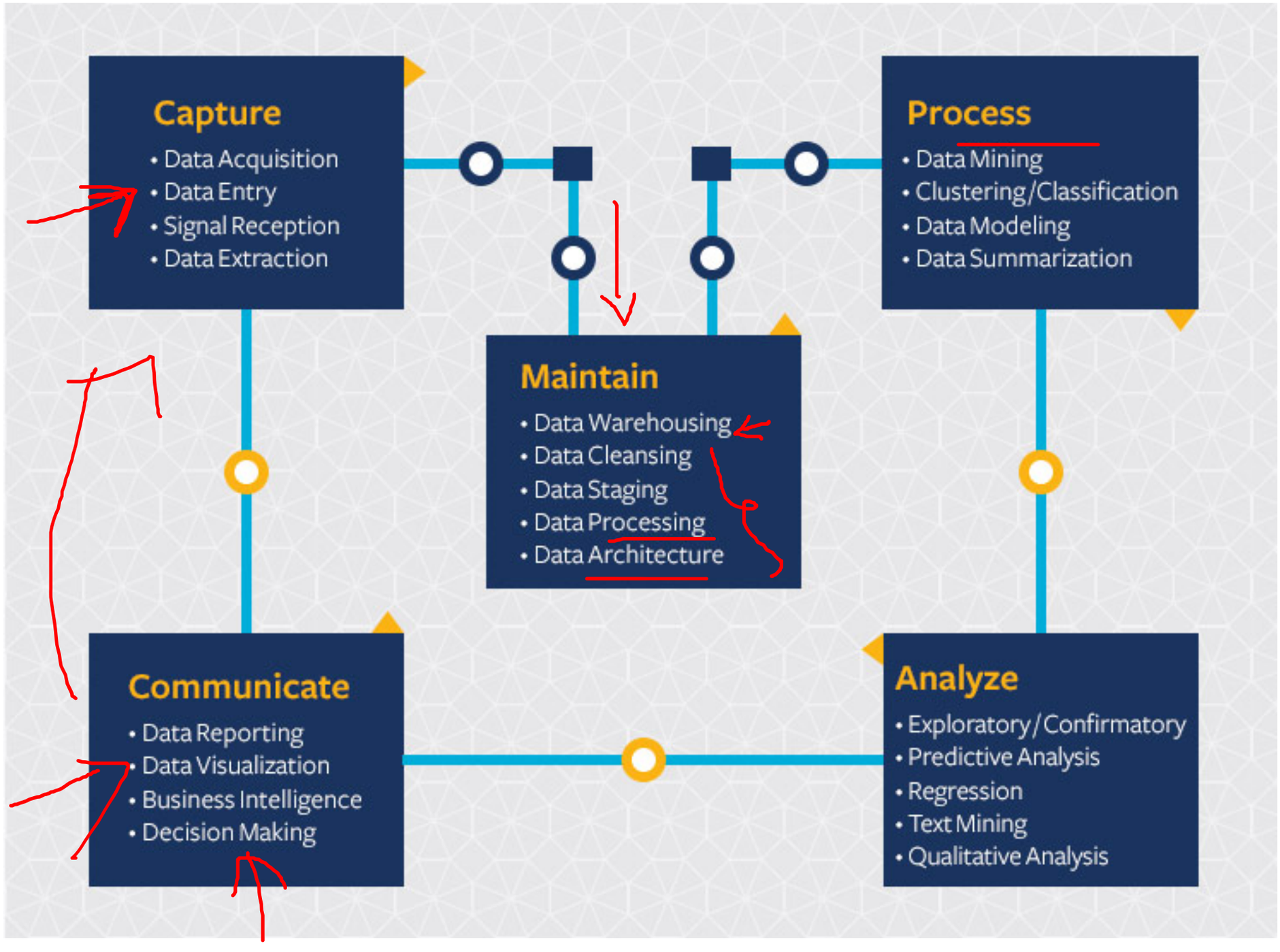
https://www.oreilly.com/library/view/doing-data-science/9781449363871/ch01.html

Data Science

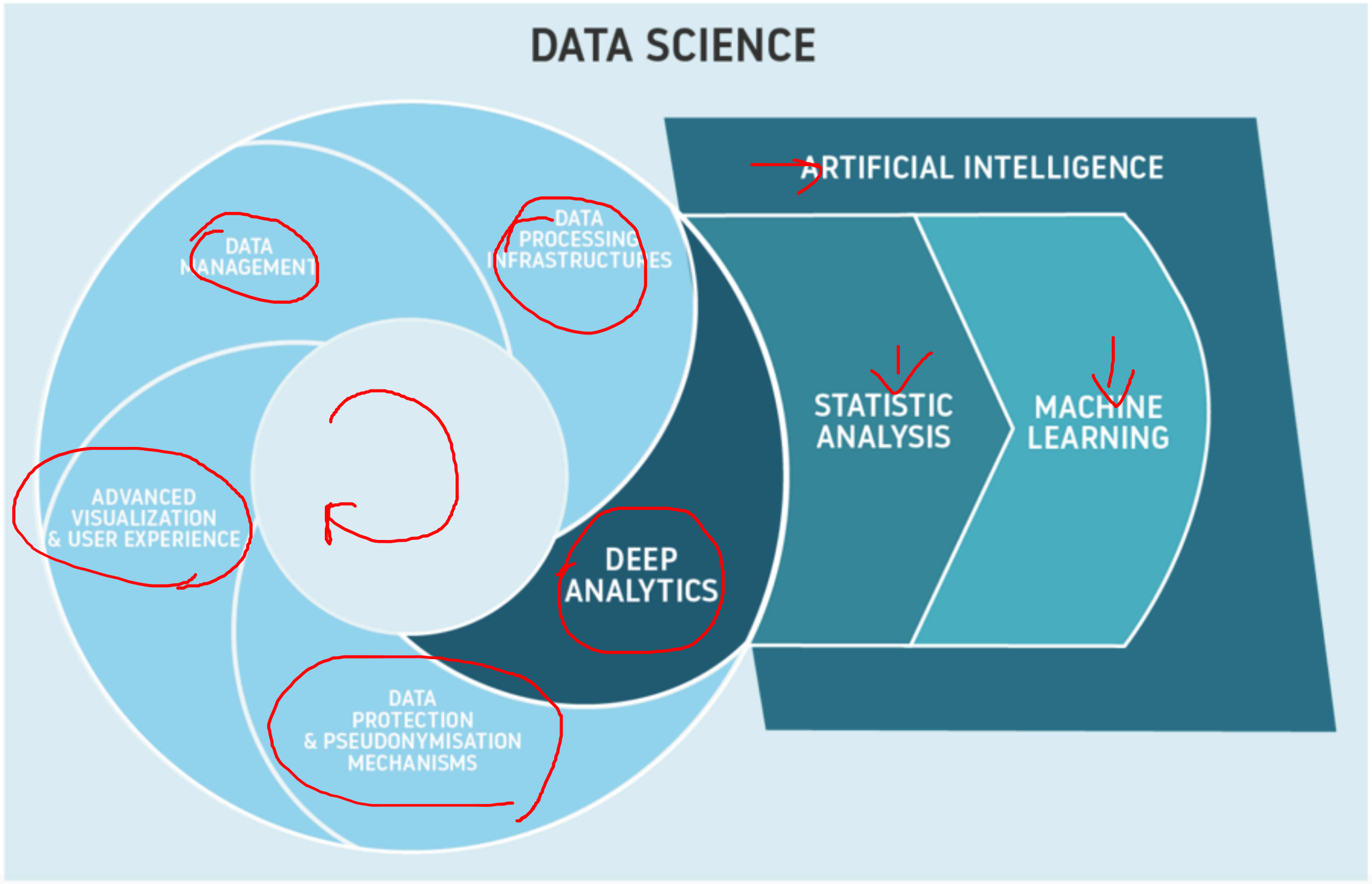
- A set of principles, problem definitions, algorithms and processes for extracting non-obvious and useful patterns from large data sets.
- Many of the elements of data Science have been developed in related fields such as machine learning and data mining.
- Data Science takes up other challenges such as capturing, cleaning and transforming unstructured social media and web data; the use of big data Technologies to store and process big unstructured data sets.

John D. Kelleher and Brendan Tierney. Data Science. The MIT Press,[2018]| Series: The MIT Press essential knowledge series| Cambridge, MA

University of Berkeley
Data Science Circle



Fonte: https://ischoolonline.berkeley.edu/data-science/what-is-data-science/



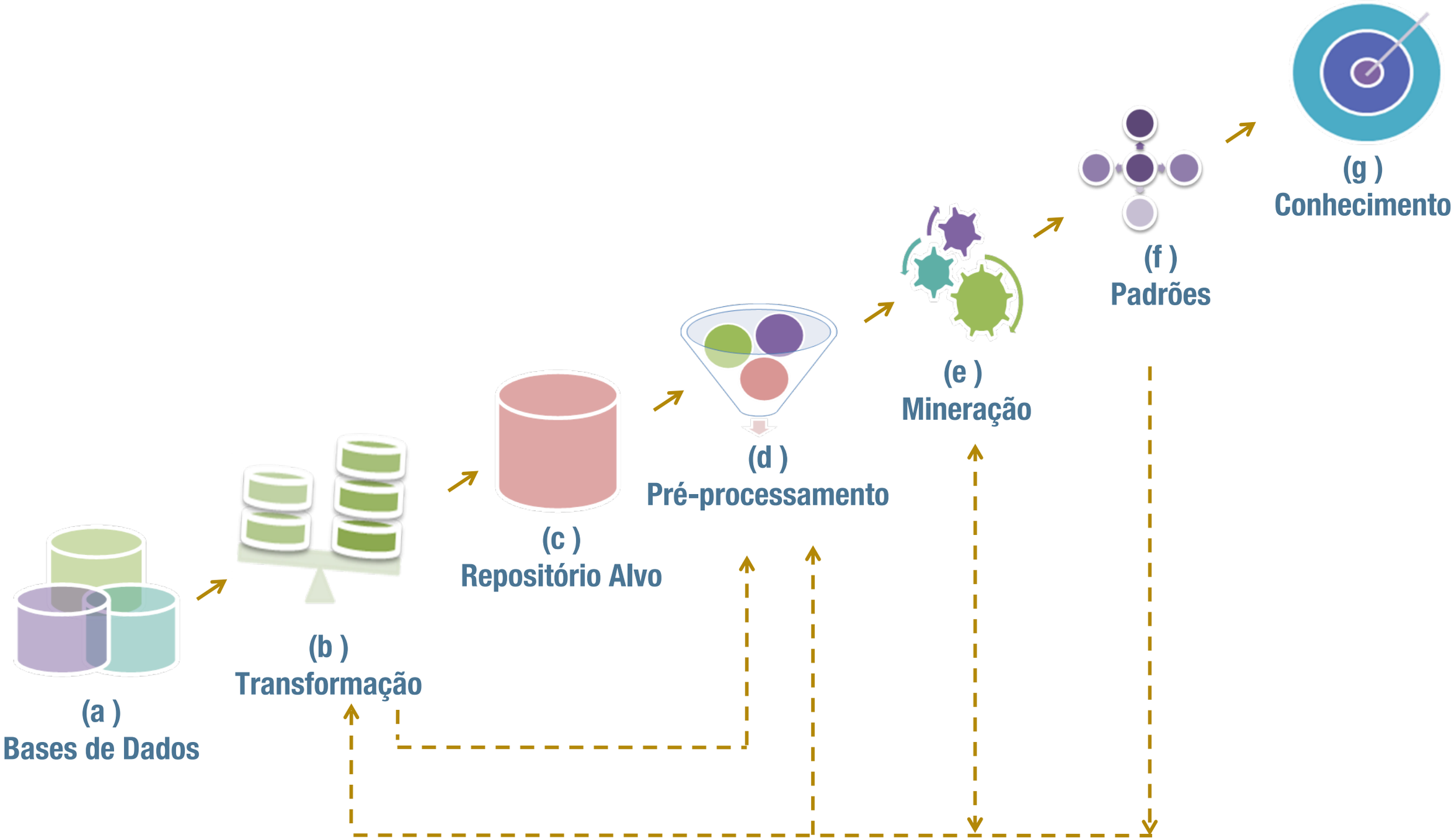
Fonte: https://datascience.aero/artificial-intelligence-vs-data-science/

KDD

- Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data.
- ...
- KDD refers to the overall pro- cess of discovering useful knowledge from da- ta, and data mining refers to a particular step in this process. *Data mining* is the application of specific algorithms for extracting patterns from data.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discove.. in databases. *AI magazine*, *17*(3), 37-37.

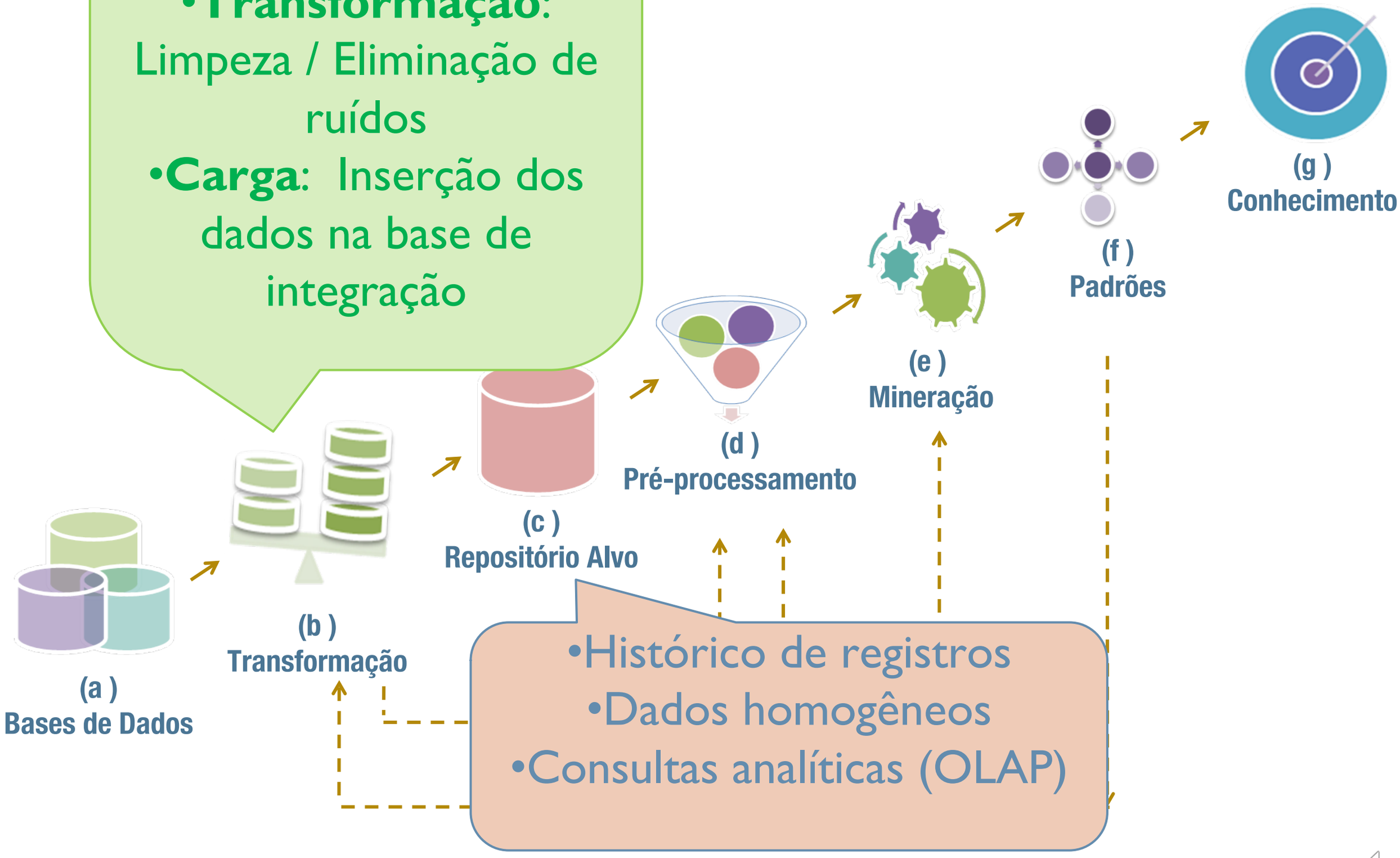
Processo de KDD



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

Pr

- **Extração**: Percorre base de dados / Extrai dados significantes
- **Transformação**: Limpeza / Eliminação de ruídos
- **Carga**: Inserção dos dados na base de integração



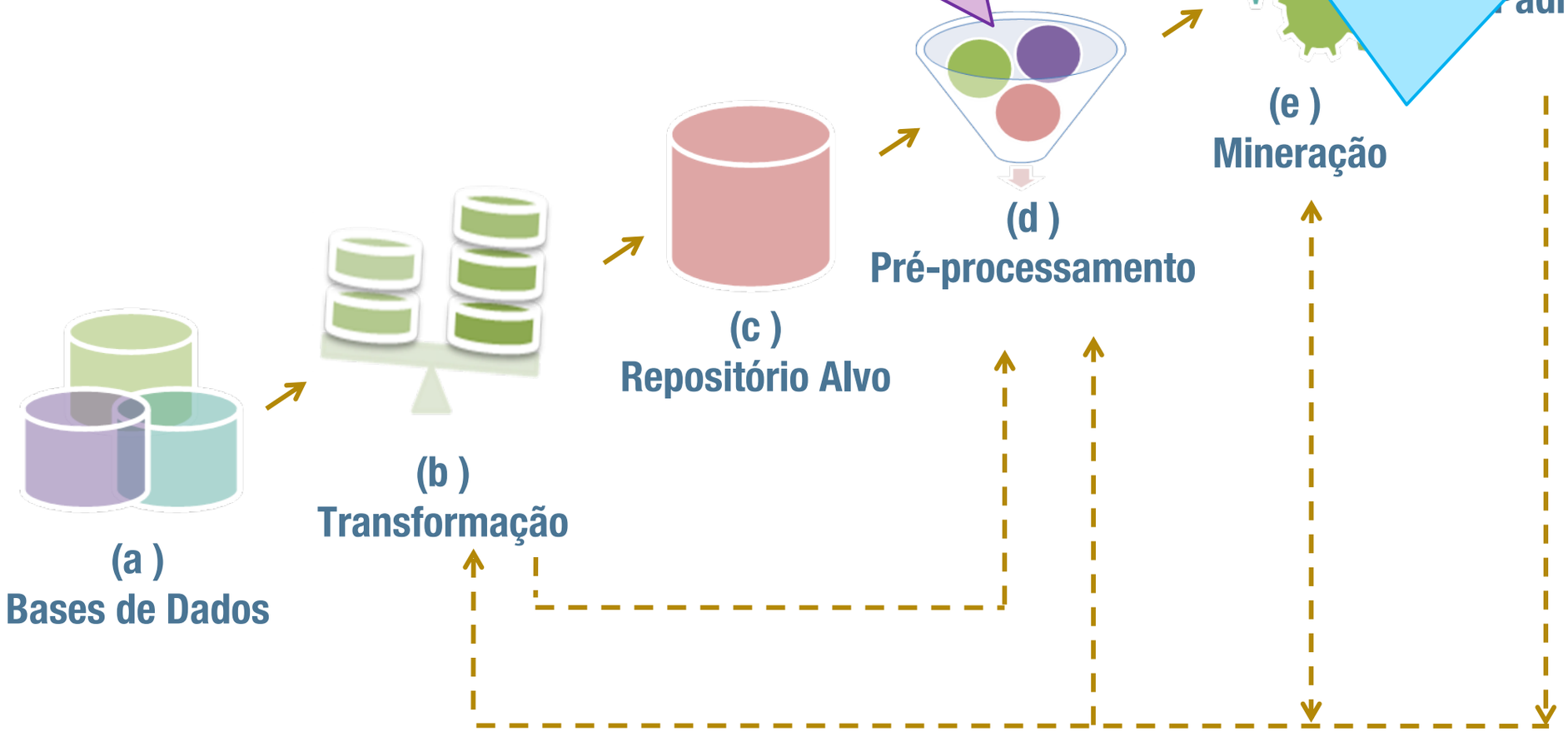
- Histórico de registros
- Dados homogêneos
- Consultas analíticas (OLAP)

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

Process

- 85% do processo de KDD
- Qualidade dos dados
- Resultados de mineração mais satisfatórios

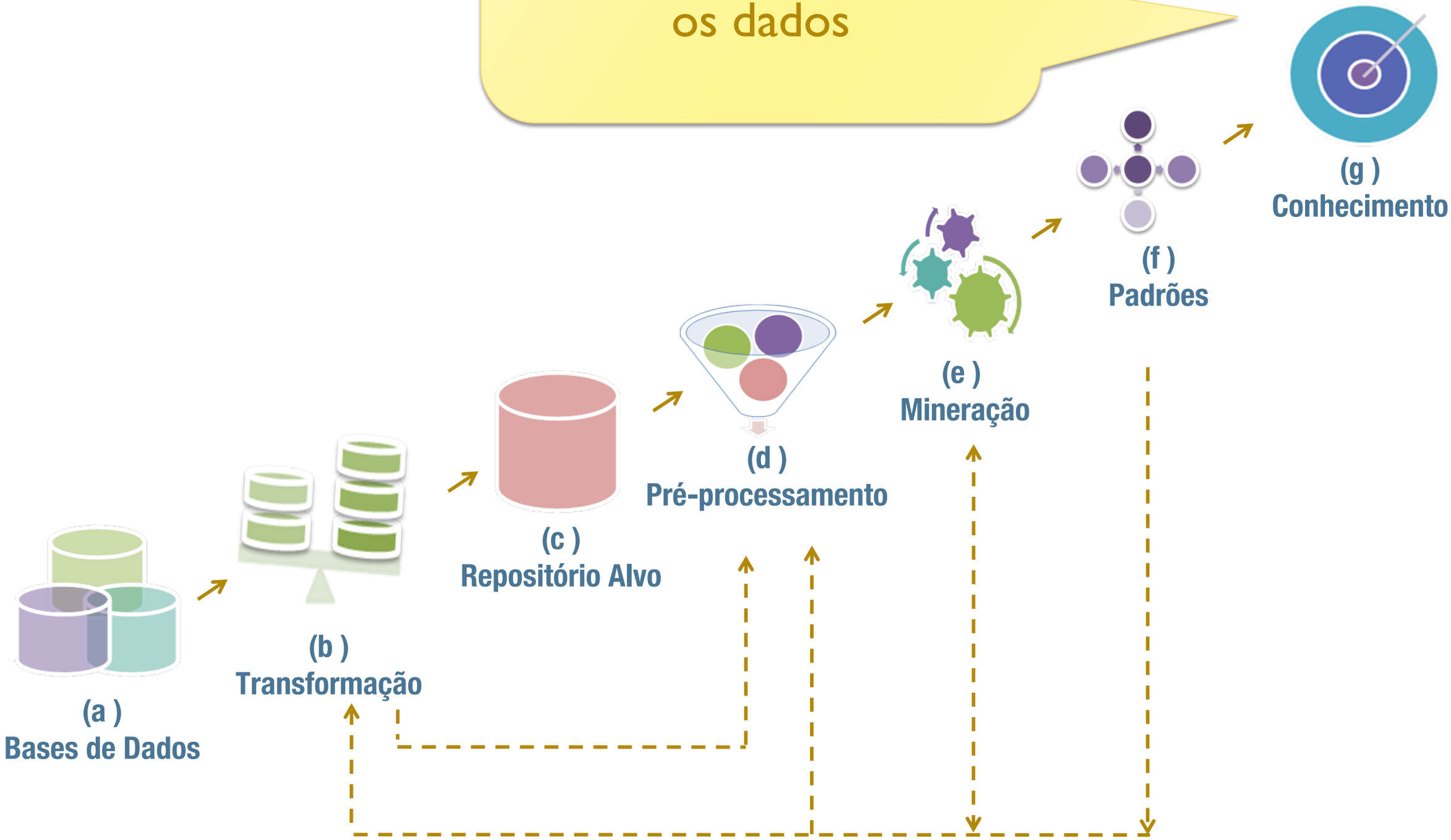
- Análise de uma série de dados
- Relacionamentos não esperados
- Resultados compreensíveis e especialmente úteis



Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, *17*(3), 37-37.

Processo de K

•Modelos induzidos
•Modelos que indicam relacionamentos entre os dados



Big Data

- Big Data concern large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data are now rapidly expanding in all science and engineering domains, including physical, biological and biomedical sciences

- HACE Theorem. Big Data starts with large-volume, heterogeneous, autonomous sources with distributed and decentralized control, and seeks to explore complex and evolving relationships among data.

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2013). Data mining with big data. *IEEE transactions on knowledge and data engineering*, *26*(1), 97-107.

IA

Machine Learning

Deep Learning

Data mining

Inteligência Artificial

- The field of **artificial intelligence,** or AI, attempts to understand intelligent entities
- Why study?
  - learn more about ourselves -> *build* intelligent entities as well as understand them
  - intelligent entities are interesting and useful in their own right

- "Although no one can predict the future in detail, it is clear that computers with human-level intelligence (or better) would have a huge impact on our everyday lives and on the future course of civilization"

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Inteligência Artificial

- It was formally initiated in 1956, when the name was coined, although at that point work had been under way for about five years
- **Definition**: Historically, all four approaches have been followed. As one might expect, a tension exists between approaches:
  - centered around humans -> empirical science, involving hypothesis and experimental
  - centered around rationality -> combination of mathematics and engineering. People in each group sometimes cast aspersions on work done in the other groups, but the truth is that each direction has yielded valuable insights

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

"The exciting new effort to make computers think . . . *machines with minds,* in the full and literal sense" (Haugeland, 1985) "[The automation of] activities that we associate with human thinking, activities such as decision-making, problem solving, learning . . ." (Bellman, 1978)	"The study of mental faculties through the use of computational models" (Charniak and McDermott, 1985) "The study of the computations that make it possible to perceive, reason, and act" (Winston, 1992)
"The art of creating machines that perform functions that require intelligence when performed by people" (Kurzweil, 1990) "The study of how to make computers do things at which, at the moment, people are better" (Rich and Knight, 1991)	"A field of study that seeks to explain and emulate intelligent behavior in terms of computational processes" (Schalkoff, 1990) "The branch of computer science that is concerned with the automation of intelligent behavior" (Luger and Stubblefield, 1993)

Figure 1.1 Some definitions of AI. They are organized into four categories:

Systems that think like humans.	Systems that think rationally.
Systems that act like humans.	Systems that act rationally.

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Inteligência Artificial – Algoritmos/técnicas

- Agentes inteligentes (sistemas multiagentes)
- Algoritmos de busca
- Planejamento (planning)
- Raciocínio probabilístico – redes bayseanas
- Tomadas de decisões simples e complexas
- Aprendizagem: por exemplos, de modelos probabilísticos, por reforço
- Processamento da linguagem natural
- Robótica

Russell, S., & Norvig, P. (2002). Artificial intelligence: a modern approach.

Machine learning – aprendizado de máquina

- Área de estudo que fornece aos computadores a habilidade de aprender sem serem explicitamente programados [Arthur Samuel (1959)].

- O campo do aprendizado de máquina está preocupado com a questão de como construir programas de computador que melhoram automaticamente com a experiência (Mitchell, 1997)
  - A computer program is said to learn from experience A with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E

Mitchell, T. M. (1997). Machine learning.

Machine learning – aprendizado de máquina

- *Aprendizado supervisionado*
- *Aprendizado não supervisionado*
- *Aprendizado por reforço*: Um programa de computador interage com um ambiente dinâmico, em que o programa deve desempenhar determinado objetivo (por exemplo, dirigir um veículo). É fornecido, ao programa, feedback quanto a premiações e punições, na medida em que é navegado o espaço do problema.

Mitchell, T. M. (1997). Machine learning.

Machine learning – aprendizado de máquina

- ML focuses on the design and evaluation of algorithms for extracting patterns from data.

- Sugestão de vídeo:
- https://www.youtube.com/watch?v=9QErWiClGjM

John D. Kelleher and Brendan Tierney. Data Science. The MIT Press,[2018]| Series: The MIT Press essential knowledge series| Cambridge, MA

Machine learning

- Descoberta de uma hipótese na forma de uma regra para definir que clientes de um supermercado devem receber material de propaganda de um novo produto ➔ usar informações de compras passadas dos clientes cadastrados

- **PROCESSO DE INDUÇÃO DE UMA HIPÓTESE A PARTIR DA EXPERIENCIA PASSADA ➔ APRENDIZADO DE MÁQUINA**

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Data mining

Mineração de dados diz respeito à extração não-trivial de informação implícita, previamente desconhecida e potencialmente útil em grandes conjuntos de dados.

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Data mining

Many people treat data mining as a synonym for another popularly used term, **knowledge discovery from data**, or **KDD**, while others view data mining as merely an essential step in the process of knowledge discovery. (Livro Han)

Data mining as a step in the process of knowledge discovery.

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Data mining

Data mining: an essential process where intelligent methods are applied to extract data patterns

Data mining is the process of discovering interesting patterns and knowledge from large amounts of data. The data sources can include databases, data warehouses, the Web, other information repositories, or data that are streamed into the system dynamically.

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Juntando tudo...

Applications of DS, ML and AL

zeolearn

Data Science

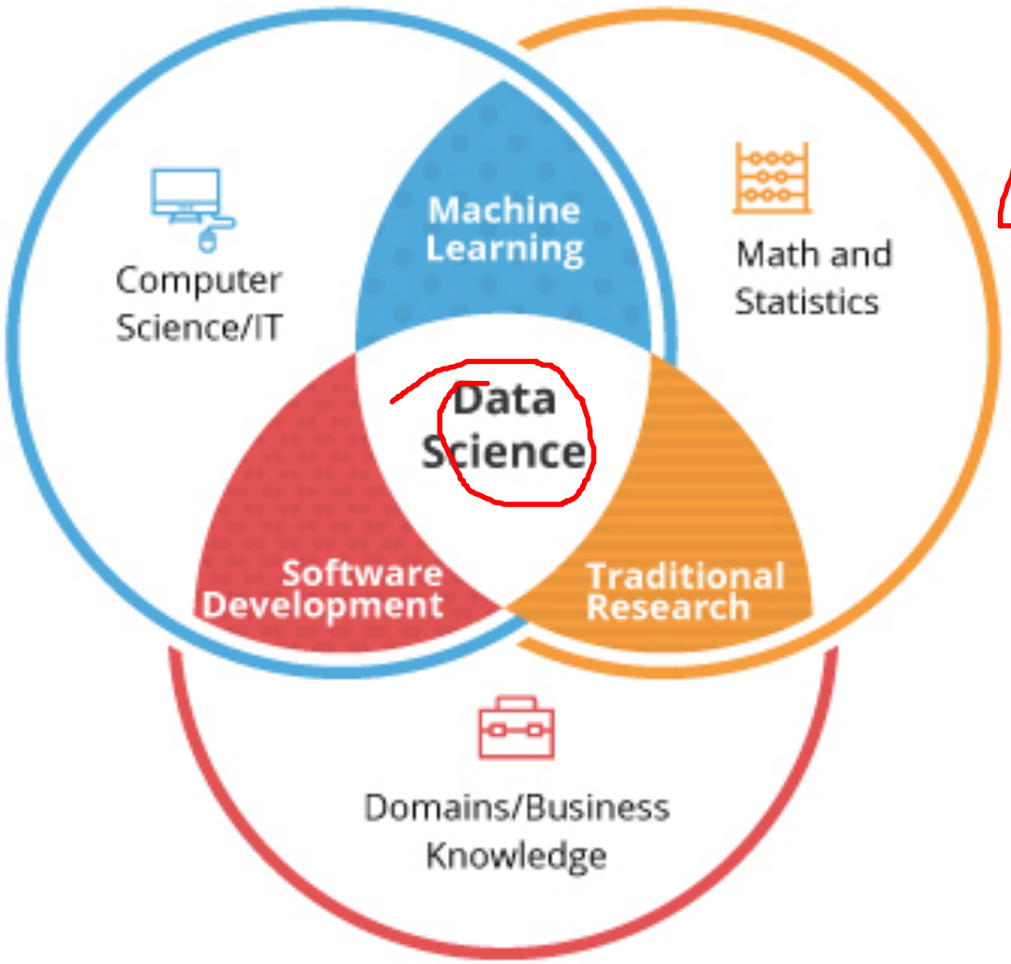
a. Time Series Forecasting
b. Recommendation Engines

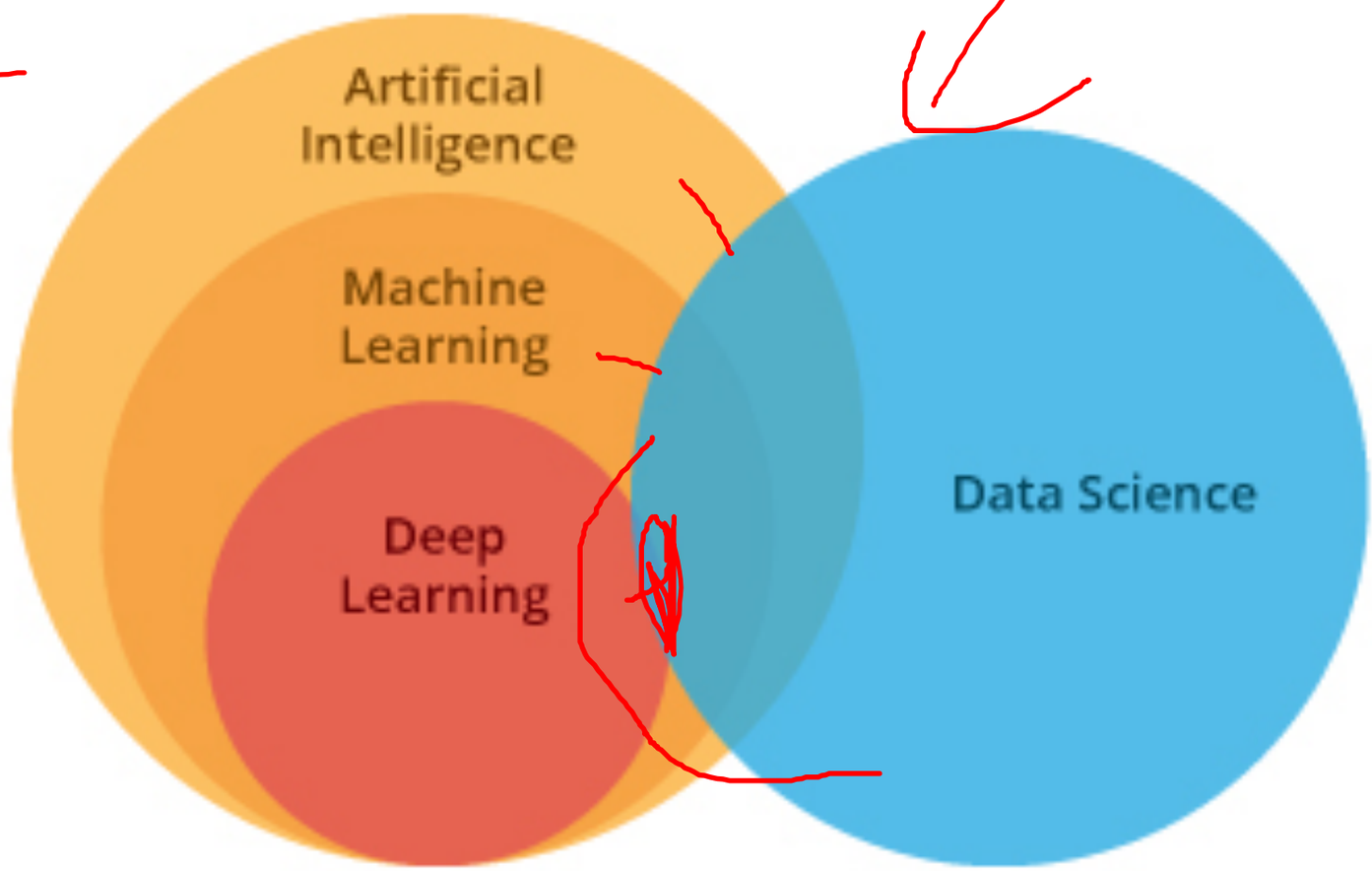
Machine Learning

a. Computer Vision
b. Natural Language Processing

Artificial Intelligence

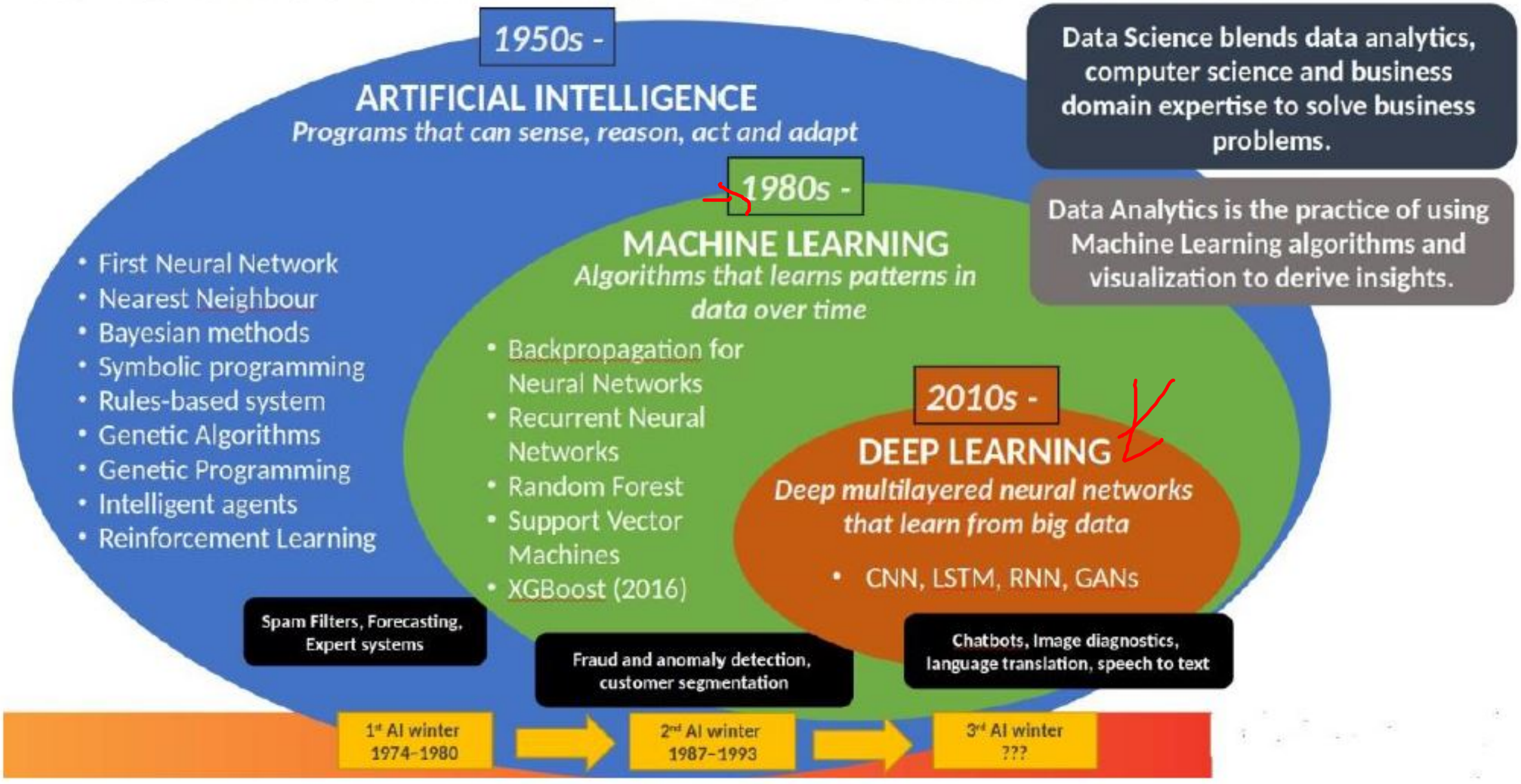
a. Game AI
b. Search
c. Robotics



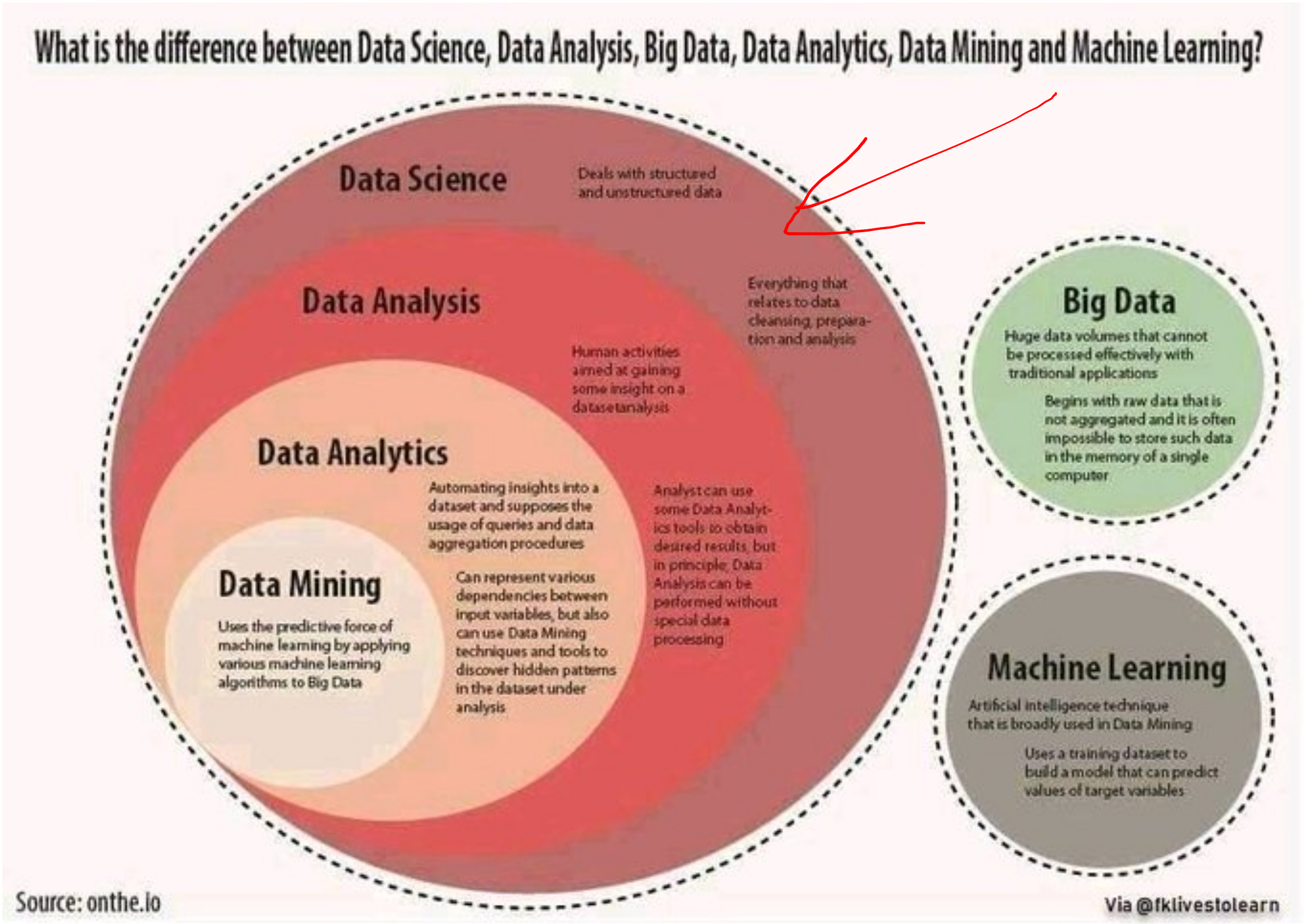


Fonte: https://www.zeolearn.com/magazine/data-science-vs-machine-learning-artificial-intelligence

AI VS ML VS DATA ANALYTICS VS DATA SCIENCE



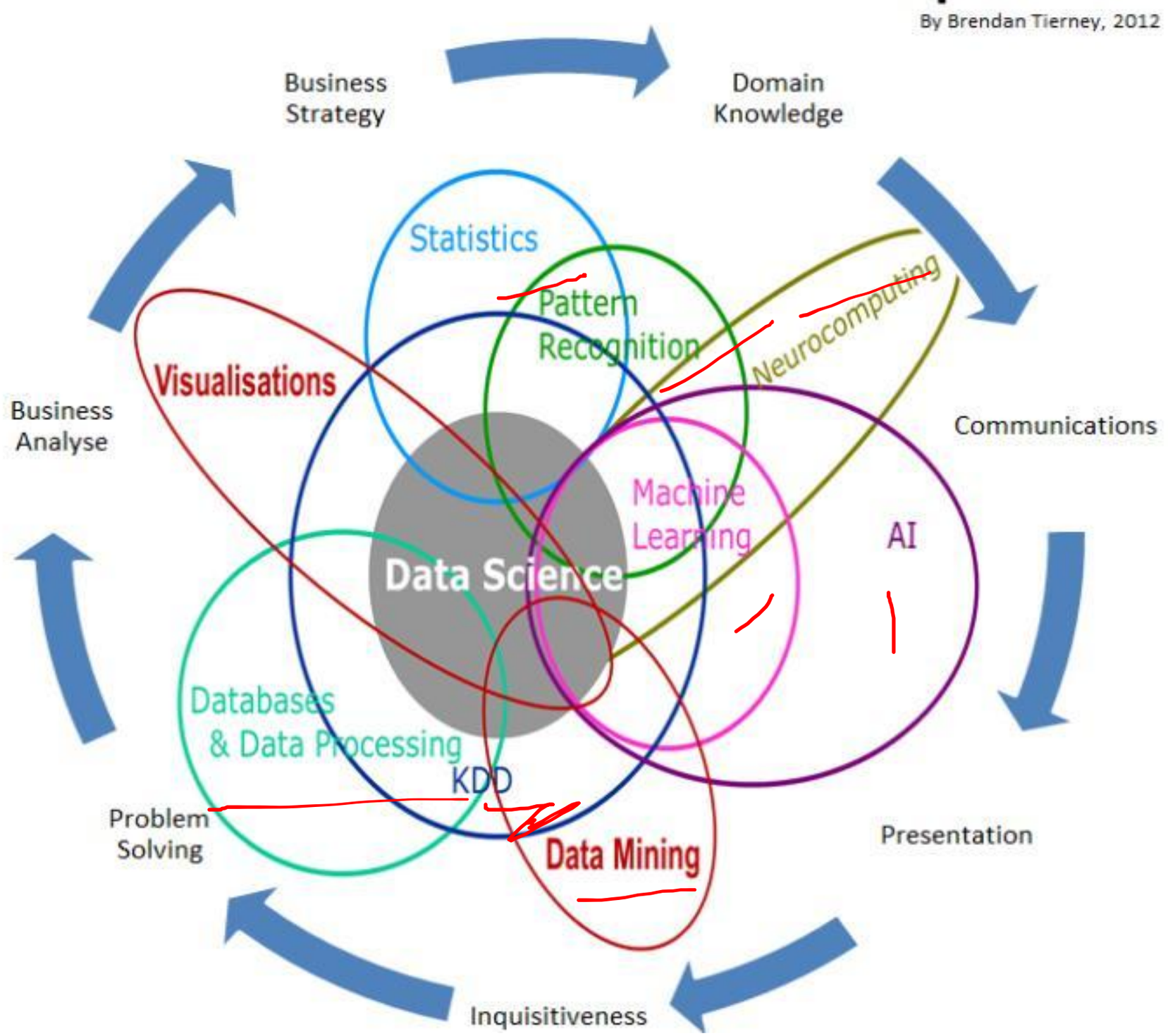
Fonte: https://www.kaggle.com/getting-started/127645



Fonte: https://medium.com/technicity/design-thinking-humanizes-data-science-more-5a666119c8b1

Fonte: https://br.pinterest.com/pin/607704543450983203/

Data Science Is Multidisciplinary



Fonte: https://www.datasciencecentral.com/profiles/blogs/difference-of-data-science-machine-learning-and-data-mining

Proposta de atividade

- Faça um infográfico relacionando os termos (não limitado):
  - Data Science
  - Data mining
  - Big Data
  - IA
  - Machine Learning
  - Deep Learning
  - Statistics
  - Supervised and unsupervised learning
  - ...

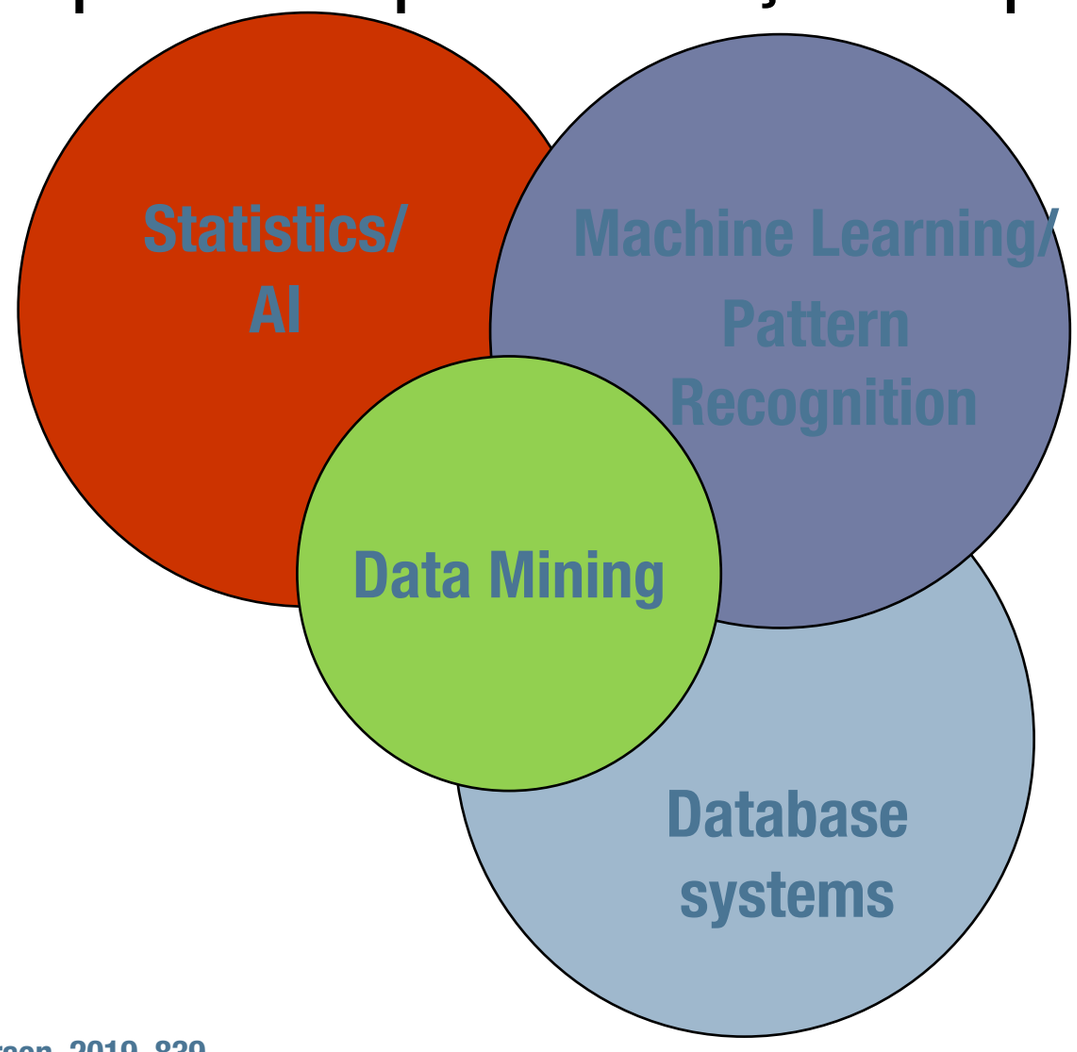
Nossa disciplina

- Vamos tratar data Science/KDD de forma similar como o processo completo de descoberta de conhecimento

- Vamos apresentar as técnicas sob um ponto de vista de Data mining, porém é similar a organização dos métodos de aprendizado

Mineração de Dados

- É uma das etapas do processo de KDD que consiste na aplicação de algoritmos específicos para extração de padrões (modelos) dos dados

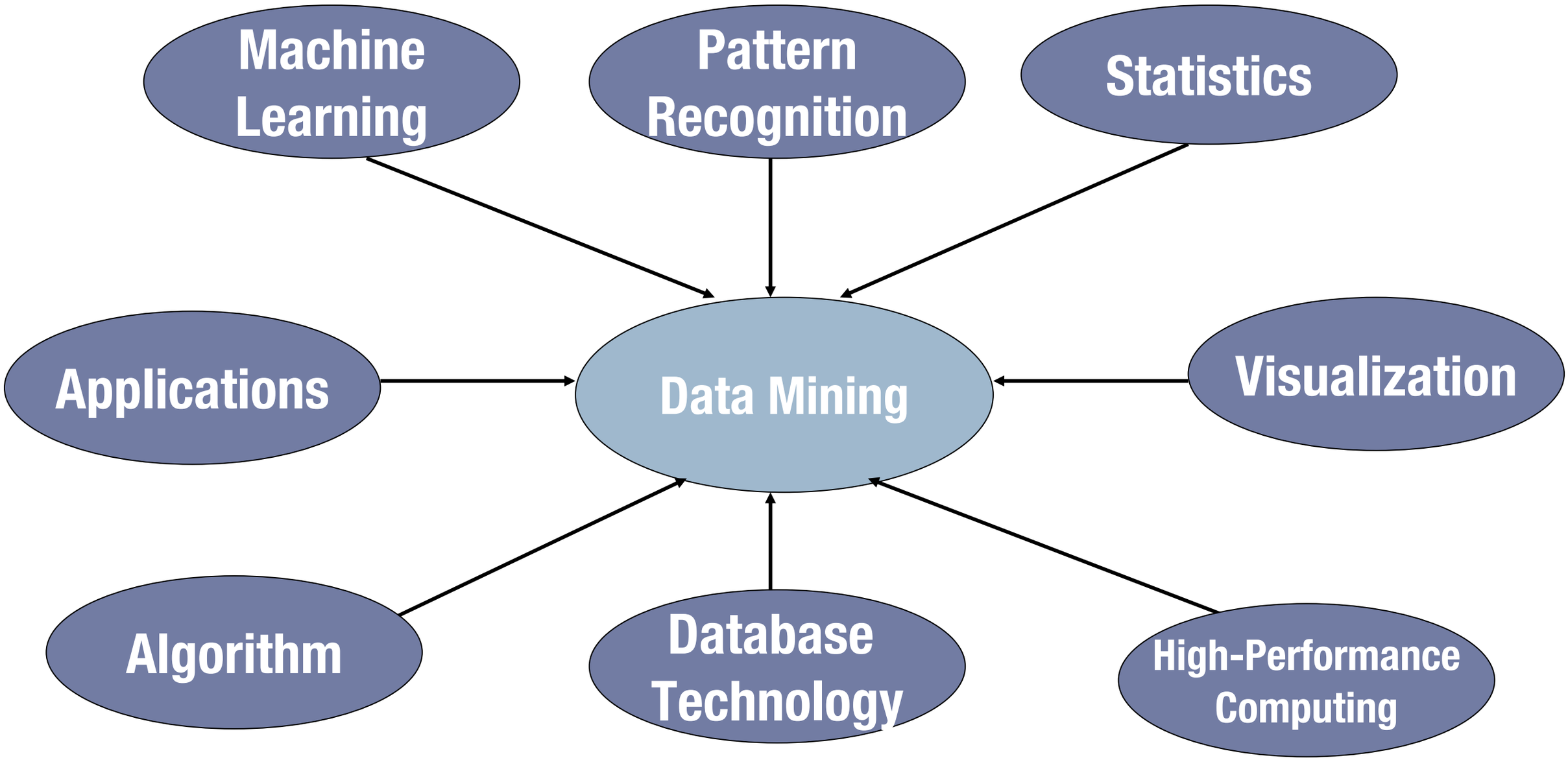


TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Mineração de Dados



TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Tarefas de Aprendizado

- Tarefas Descritivas
  - Relação entre dados

→ Associação

→ Agrupamento

- Tarefas Preditivas
  - Previsão sobre dados

→ Classificação

→ Regressão

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Tarefas de Aprendizado

- Tarefas Descritivas
  - Relação entre dados

→ Associação

→ Agrupamento

META: EXPLORAR OU DESCREVER UM CONJUNTO DE DADOS.

PARADIGMA NÃO SUPERVISIONADO : Não há um atributo alvo (ou de saída)

META: Encontrar uma função (modelo) a partir dos dados de treinamento que possa ser utilizada para prever um rótulo ou valor que caracterize um novo exemplo, com base nos valores de seus atributos de entrada.

PARADIGMA SUPERVISIONADO :PRESENÇA DE UM SUPERVISOR EXTERNO -> CONHECE O RÓTULO DOS EXEMPLOS

- Tarefas Preditivas
  - Previsão sobre dados

→Classificação

→Regressão

TAN, Pang-Ning et al. Introduction to data mining. 2nd ed. New York: Pearson, 2019. 839

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data mining: concepts and techniques. 3rd ed. Amsterdam: Elsevier, 2012. 703 p.

Hierarquia de Aprendizado

FACELI, Katti et al. Inteligência artificial: uma abordagem de aprendizado de máquina. Rio de Janeiro: LTC, 2017. 378 p.