MINERAÇÃO DE DADOS

Veronica Oliveira de Carvalho

Introdução

- Sem perceber, as pessoas geram dados a todo momento
  - Aplicam para um cartão de fidelidade
    - Empresa aérea, supermercado, etc.
  - Fazem uma compra com cartão de débito ou crédito
  - Navegam na internet
- Esses dados são armazenados em computadores (pessoais ou nuvens)

Introdução

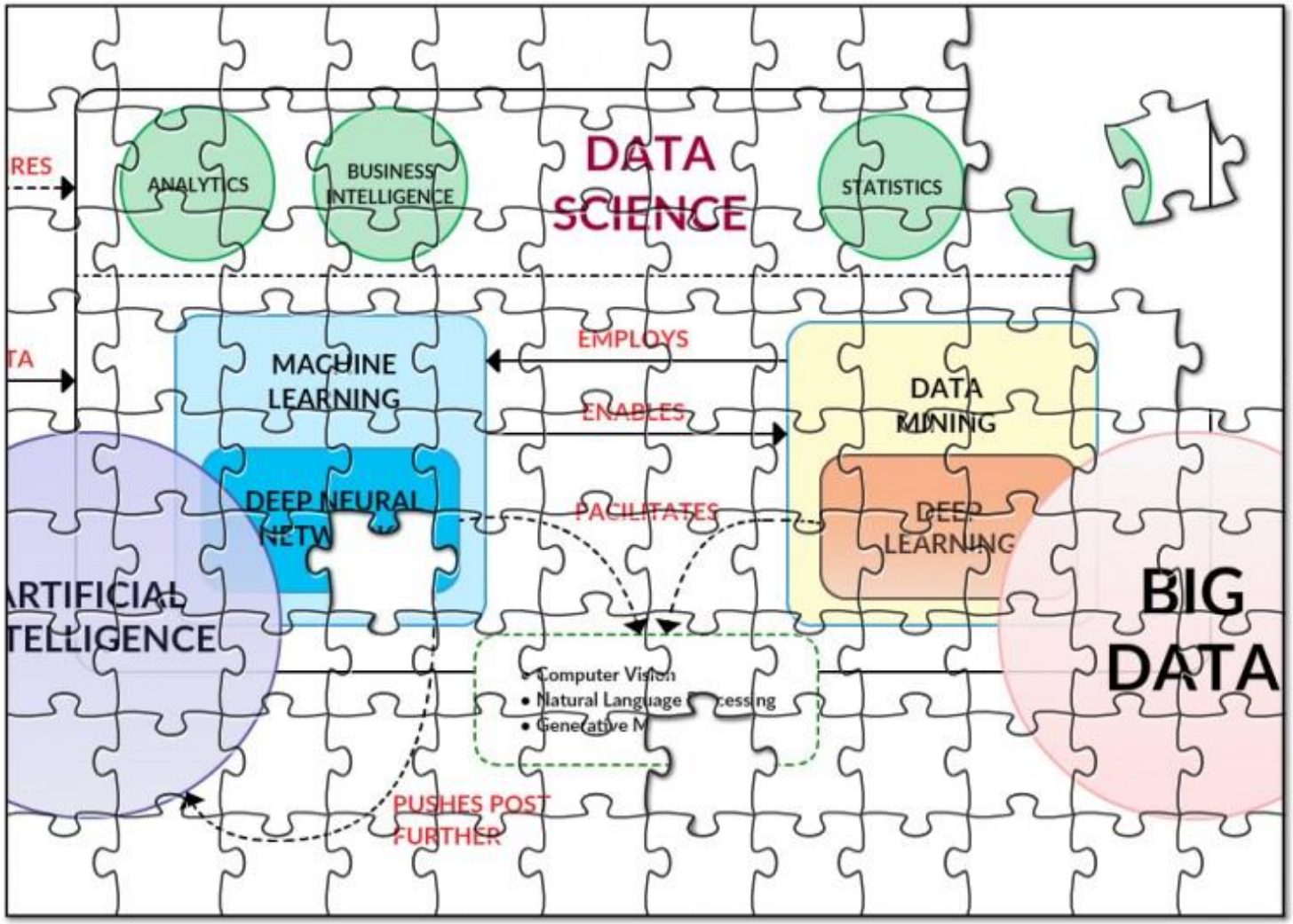
- Esses dados geralmente contêm informações relevantes
  - Uma vez analisados, podem trazer vários benefícios
- No passado,
  - Poucas empresas geravam dados
  - Todo o resto (empresas e pessoas) consumia dados
- Hoje em dia,
  - Todo mundo produz dados
  - Todo mundo consome dados

Introdução

What Happens on the Internet Every Minute (2023 Version) [Infographic]

https://www.domo.com/news/press/domo-releases-11th-annual-data-never-sleeps-report?utm_source=twitter&utm_medium=orgs&utm_campaign=TW_PRS_DNS11&utm_campid=7015w000000vnouAAA

Conceitos Gerais



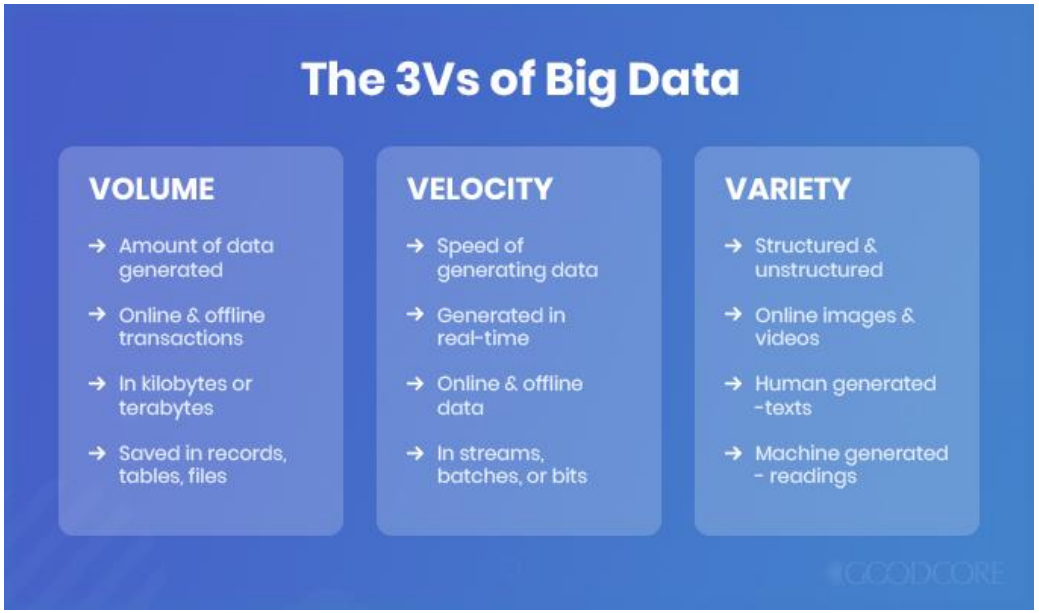
https://www.kdnuggets.com/2017/01/data-science-puzzle-revisited.html

Conceitos Gerais

- Big Data
  - Visa dar suporte à coleta e ao gerenciamento de grandes quantidades de dados
  - Permite armazenar, processar e transmitir dados cada vez maiores

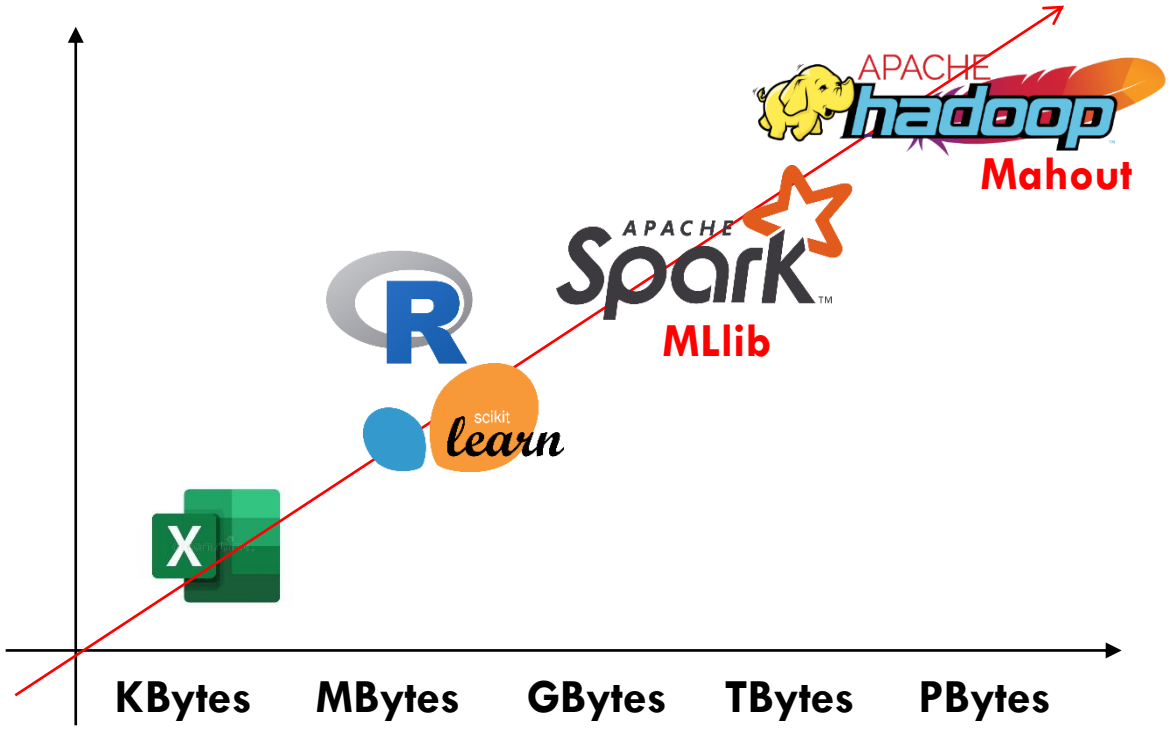
Varia de 3 a 8

https://www.goodcore.co.uk/blog/big-data-analytics/



Conceitos Gerais

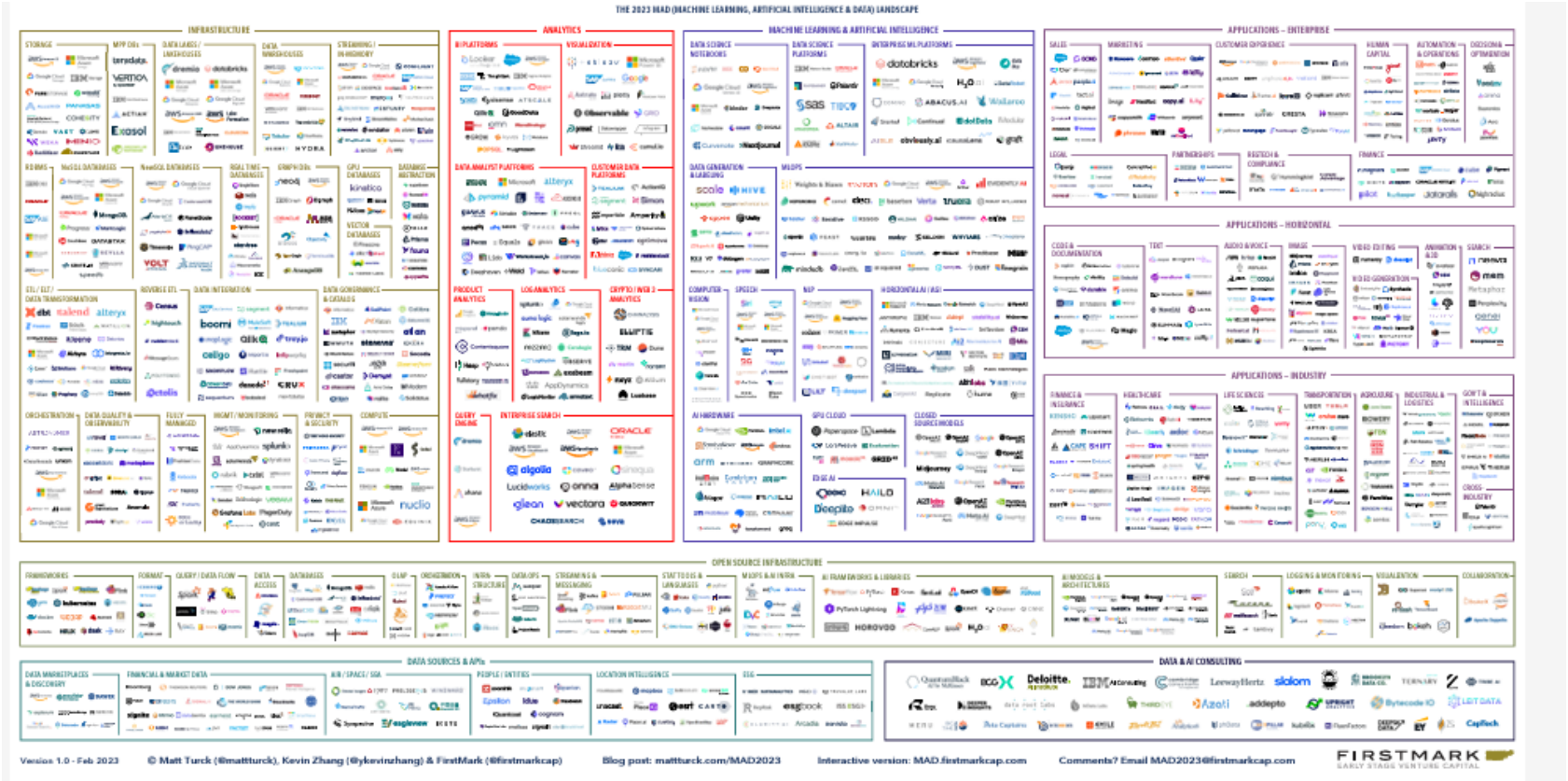
- Big Data



Conceitos Gerais

- Big Data

Panorama IA/Dados



Conceitos Gerais

- Ciência de Dados (Data Science)
  - Visa estudar princípios, métodos e sistemas computacionais para extrair conhecimento de dados (estruturados, semiestruturados e não estruturados)
    - Pergunta chave: como encontrar de maneira eficiente conhecimento (padrões) em (grandes) conjuntos (fluxos) de dados?

Conceitos Gerais

- Aprendizado de Máquina (Machine Learning)
  - Investiga técnicas capazes de aprender a resolver problemas, de maneira automática, sem intervenção humana
  - Aplicado em vários problemas reais de modelagem, tanto descritivos, quanto preditivos

Conceitos Gerais

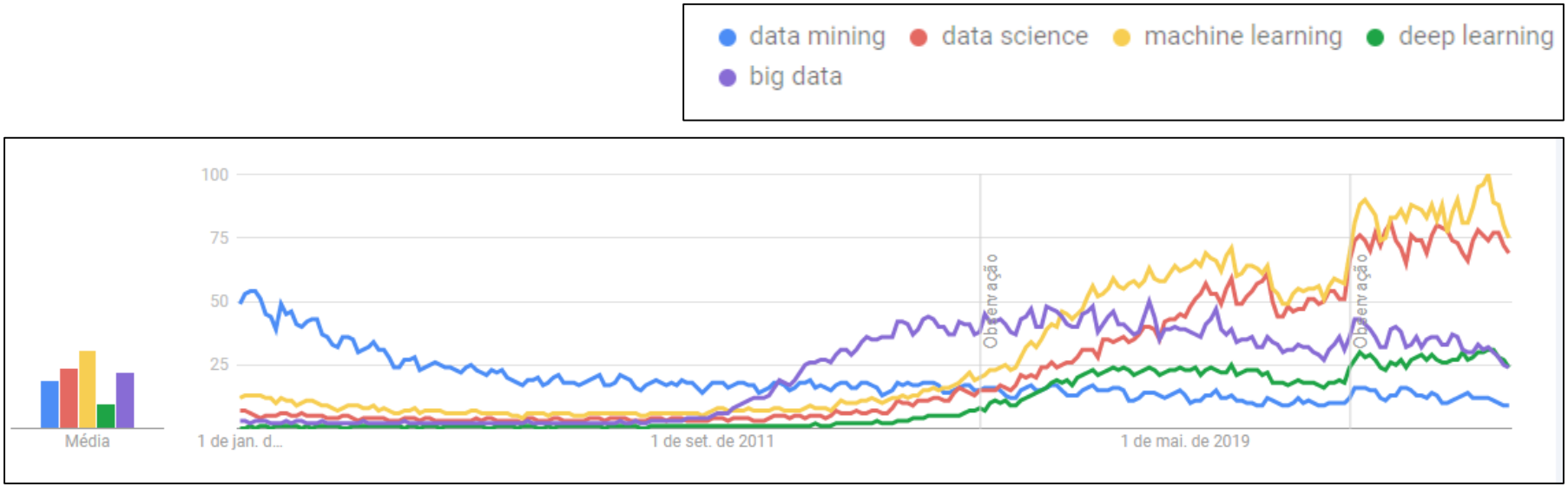
- Mineração de Dados (Data Mining)
  - A Mineração de Dados é um processo no qual algoritmos de AM são utilizados para se extrair padrões a partir de um conjunto de dados
  - O foco está na aplicação dos algoritmos, ao contrário do desenvolvimento dos algoritmos em si (AM)
  - Usada como sinônimo de KDD

Conceitos Gerais

- A Mineração de Dados é o conceito mais similar a Ciência de Dados
  - Muitas vezes vistas como sinônimos
  - Contudo, a Ciência de Dados é considerada como um superconjunto da Mineração de Dados e, portanto, um termo sucessor (uma área)

Conceitos Gerais

- Antes da explosão do termo Ciência de Dados, a Mineração de Dados teve um sucesso muito maior como um termo de pesquisa no Google



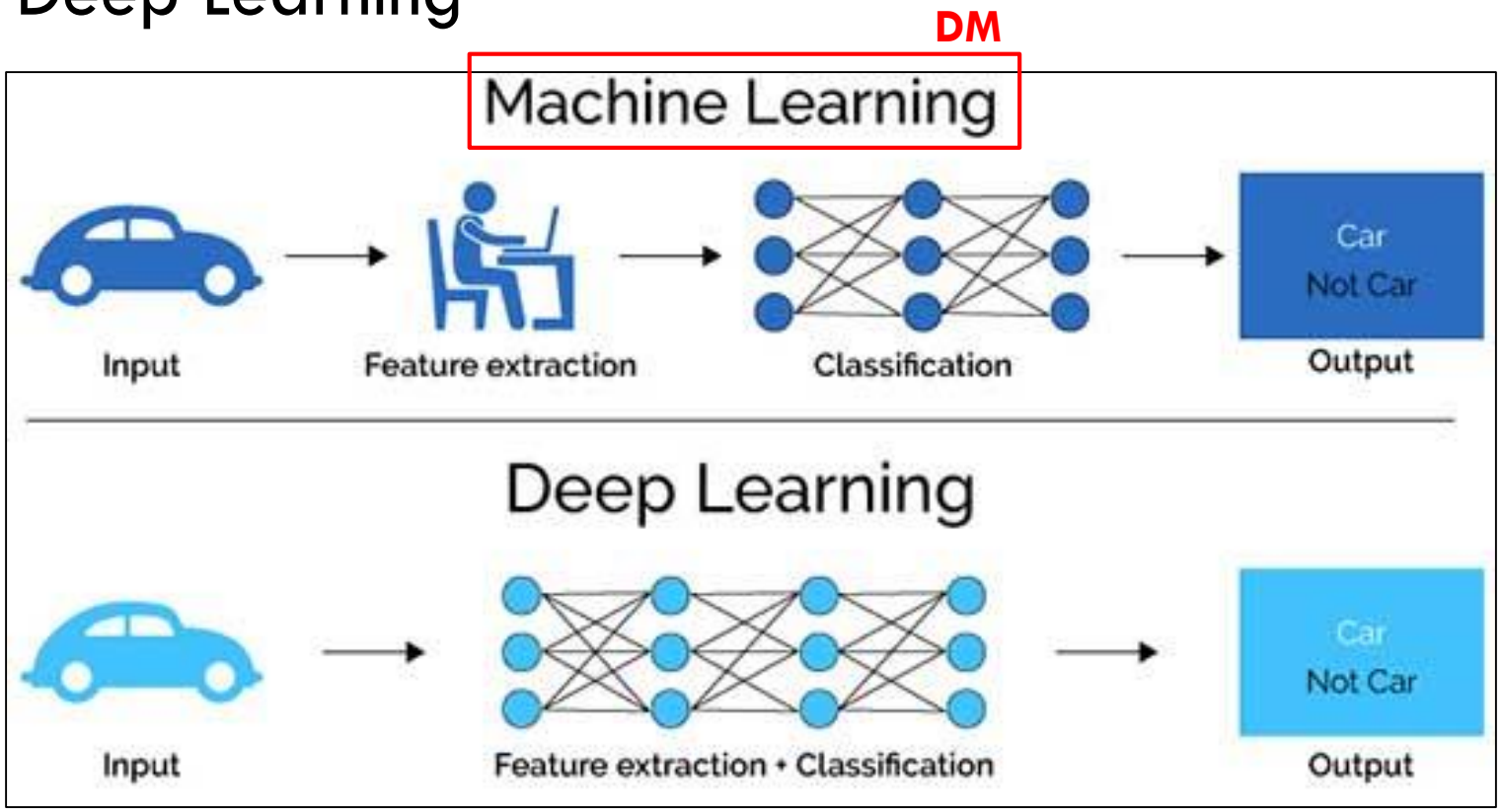
https://trends.google.com.br/trends/explore?date=all&q=data%20mining,data%20science,machine%20learning,deep%20learning,big%20data

Conceitos Gerais

- Deep Learning
  - O aprendizado profundo (Deep Learning) é o processo de aplicação de tecnologias de redes neurais profundas (Deep Neural Network), i.e., arquiteturas de redes neurais com várias camadas ocultas, para resolver problemas
  - O aprendizado profundo é um processo, como a Mineração de Dados, que emprega arquiteturas de redes neurais profundas, que são tipos específicos de algoritmos de Aprendizado de Máquina

Conceitos Gerais

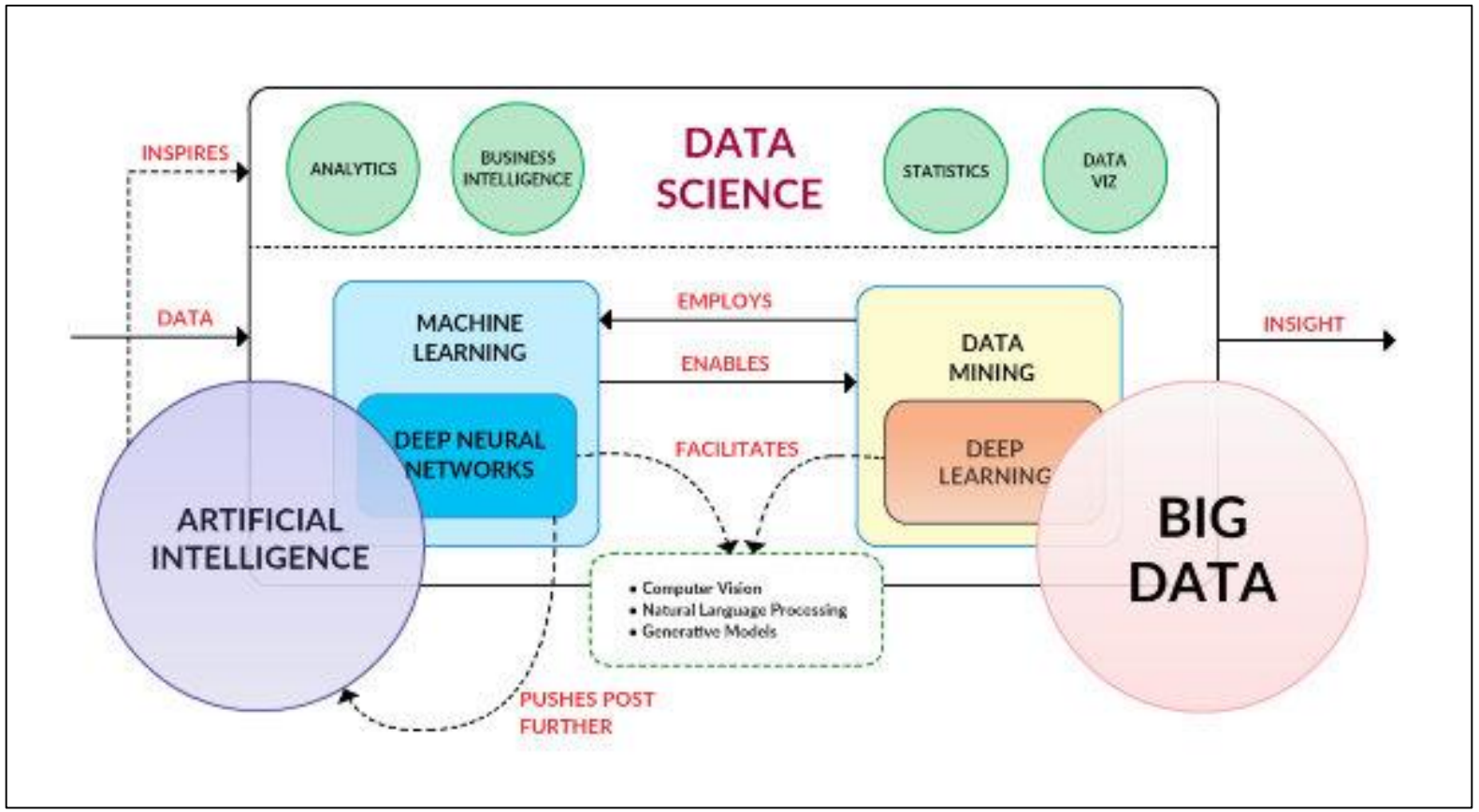
- Deep Learning



https://www.computer.org/publications/tech-news/trends/deep-learning-vs-machine-learning-whats-the-difference

Conceitos Gerais

- Lembrando que não existe consenso entre os termos



https://www.kdnuggets.com/2016/03/data-science-puzzle-explained.html

Metodologias

- A Ciência de Dados visa a extração de conhecimento. Para tanto, é necessária a utilização de metodologias para obtê-lo
  - KDD: Knowledge Discovery in Databases
  - CRISP-DM: Cross Industry Standard Process for Data Mining
  - SEMMA: Sample, Explore, Modify, Model and Assess (SAS)

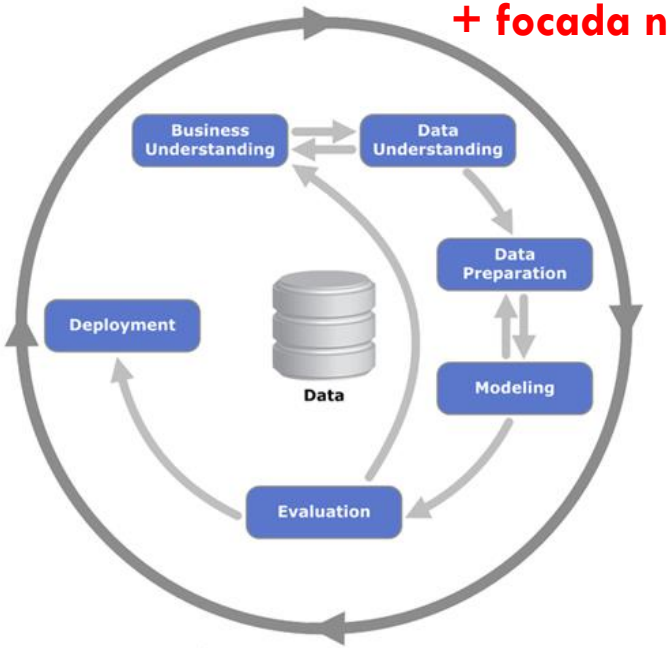
Metodologias

From Data to Insight

Cross Industry Standard Process for Data Mining (CRISP-DM)

(IBM, Teradata, Daimler AG, NCR Corporation and OHRA)

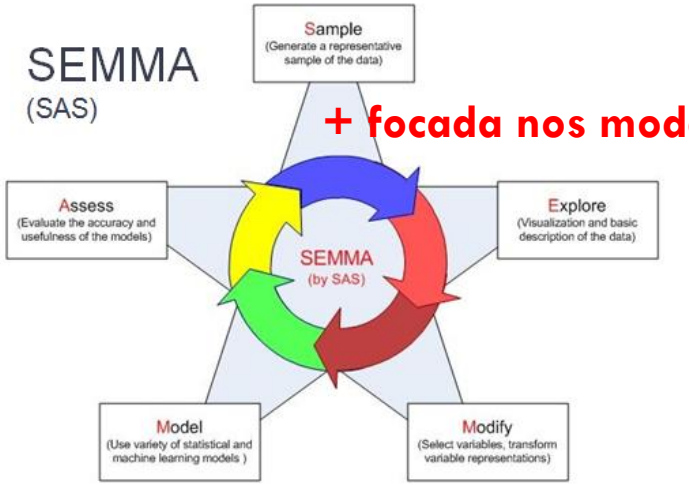
+ focada no negócio



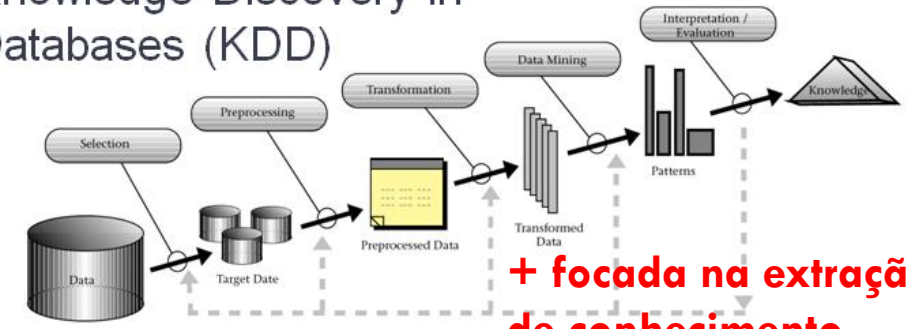
SEMMA

(SAS)

+ focada nos modelos



Knowledge Discovery in Databases (KDD)



+ focada na extração de conhecimento

BUSINESS BROADWAY

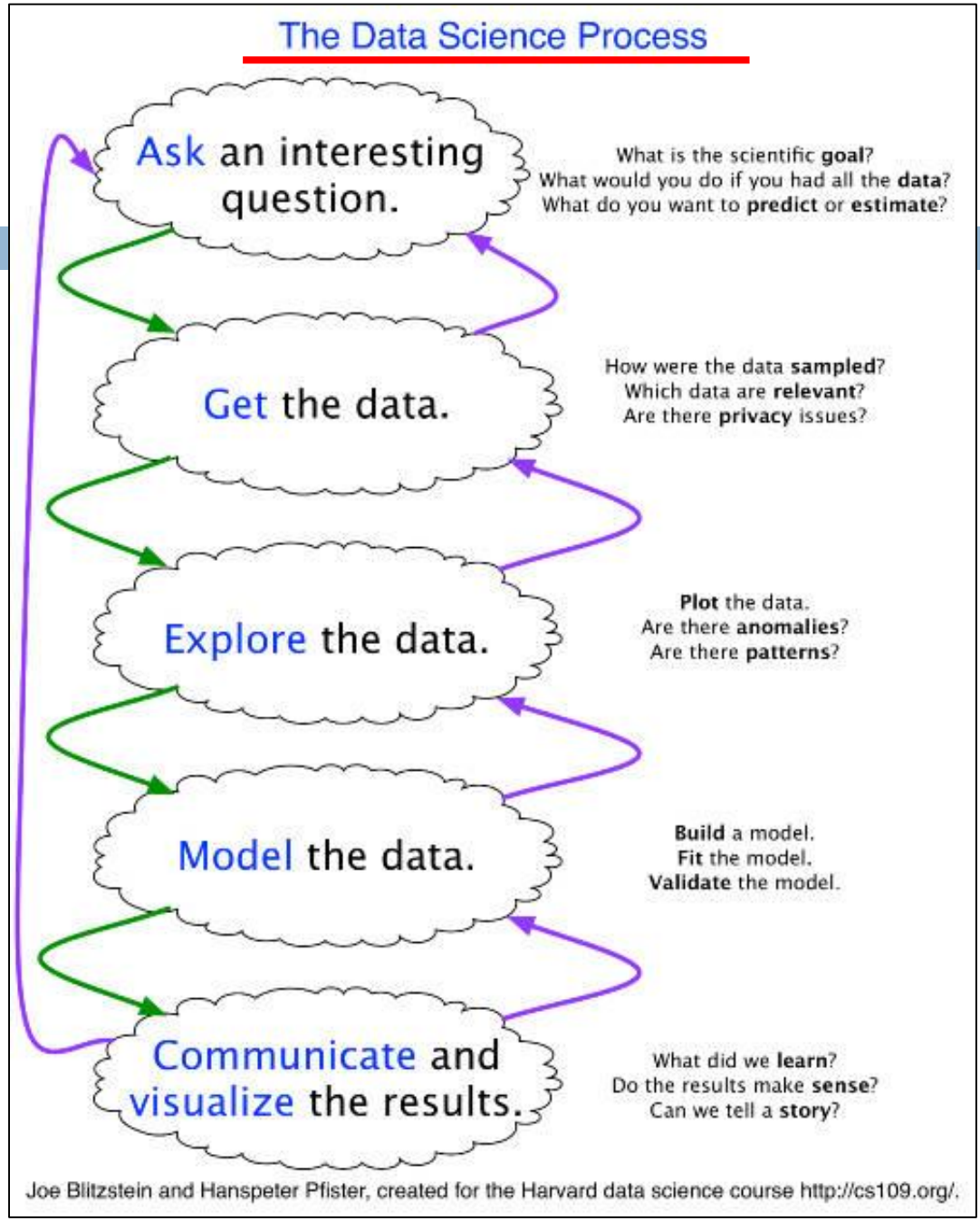
For more information on these methods, see: https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining; https://en.wikipedia.org/wiki/SEMMA; https://en.wikipedia.org/wiki/Data_mining

http://businessoverbroadway.com.convey.pro/I/0X5obVL

Metodologias

- Research Advances in Data Mining Techniques and Applications, 2024 [book]
- A survey of data mining methodologies in the environment of IoT and its variants, 2024
- Adaptations of data mining methodologies: a systematic literature review, 2020
- Data Mining Methodologies in the Banking Domain: A Systematic Literature Review, 2019 – As vezes adaptadas para domínios específicos
- KDD, SEMMA and CRISP-DM: A Parallel Overview, 2008

Metodologias

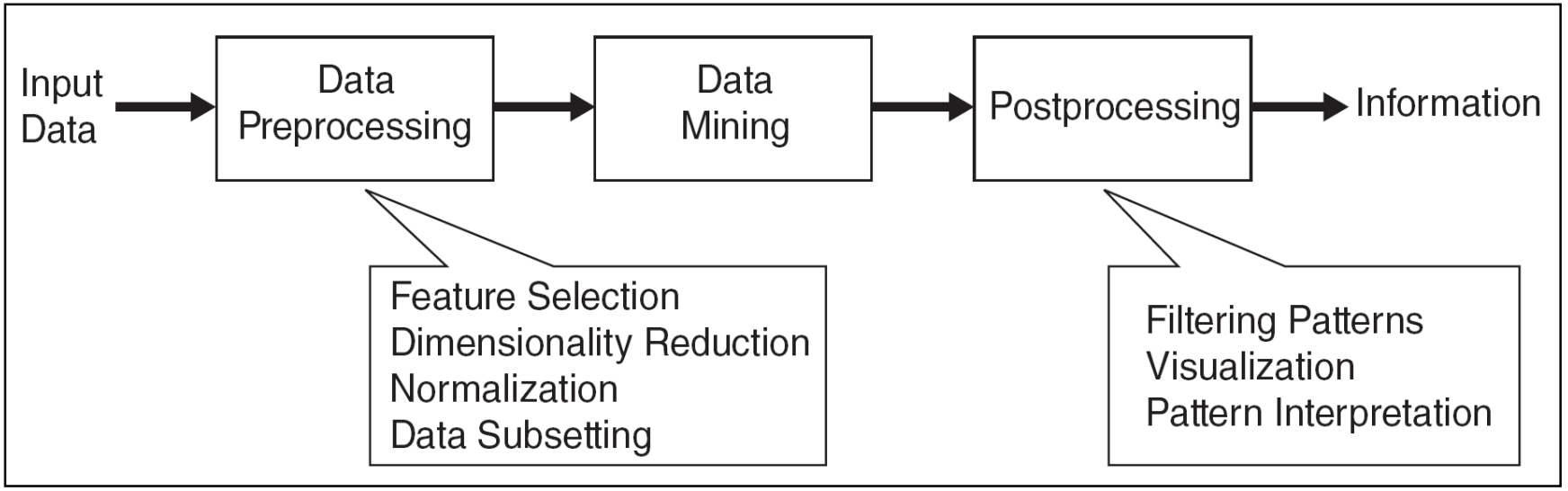


https://www.kdnuggets.com/2016/03/data-science-process-rediscovered.html

Metodologias

- The Data Science Process [https://www.kdnuggets.com/2016/03/data-science-process.html]
- Data Science Methodology for Cybersecurity Projects, 2018 – As vezes adaptadas para domínios específicos
  - KDD
  - CRISP-DM
  - FMDS (IBM) [https://www.ibm.com/downloads/cas/B1WQ0GM2]
  - TDSP (Microsoft) [https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/overview]

Metodologias



Introduction to Data Mining, 2nd Edition [Cap. 1]
Tan, Steinbach & Kumar, 2019

Recursos

- KDnuggets [https://www.kdnuggets.com/]
  - Um dos maiores sites em IA, Analytics, Big Data, Data Mining, Data Science e Machine Learning e é editado por Gregory Piatetsky-Shapiro e Matthew Mayo [https://www.kdnuggets.com/about/index.html]
  - Bases de Dados: http://www.kdnuggets.com/datasets/index.html
- Towards Data Science [https://towardsdatascience.com/]
- Machine Learning Mastery [https://machinelearningmastery.com/]

Recursos

- Repositório de bases de dados
  - UCI [http://archive.ics.uci.edu/ml/]
  - KEEL [https://sci2s.ugr.es/keel/datasets.php]
- Kaggle [http://www.kaggle.com/]
  - Competições, Datasets, Cursos, etc.

Linguagens/Ferramentas/Bibliotecas

- Java
  - Weka [https://www.cs.waikato.ac.nz/ml/weka/]
  - KEEL [http://www.keel.es/]
  - SPMF [https://www.philippe-fournier-viger.com/spmf/]
- R [pacotes]
  - RDataMining.com: R and Data Mining [http://www.rdatamining.com/]
- Python [bibliotecas]
  - Scikit-learn [https://scikit-learn.org/stable/]

Links

- https://edisciplinas.usp.br/pluginfile.php/794354/course/section/241802/talk-ICMC2016-Andre.pdf