

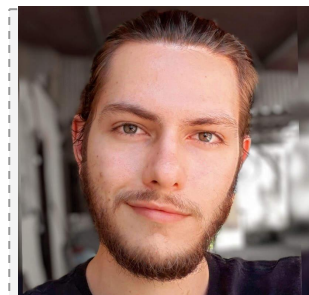


Contornos da ciência de dados: uma meta-revisão conceitual

Daniel Brito dos Santos (danielbrito@pq.uenf.br);

Annabell Del Real Tamariz (annabell@uenf.br)

UNEF – CCT – LCMAT – Ciência da Computação



Introdução

Ciência de dados é parte essencial de muitas **indústrias**, principal produto das **empresas** mais valiosas do mundo, e uma das principais ferramentas de **pesquisa científica**.



Portanto, **expandir** e **aprofundar** a ciência de dados se reflete na **ampliação** das possibilidades e efetividade de **negócios**, **pesquisas**, **tomadas de decisão** e **organização social**.



Entretanto, alguns desafios permanecem centrais à disciplina:

- formação de mão-de-obra qualificada,
- padrões de projeto,
- comunicação entre profissionais de diferentes áreas, e até mesmo
- consenso quanto a sua própria delimitação.

Nesse sentido, é interessante compreender os esforços atuais na direção de estruturar a ciência de dados como campo de pesquisa e prática.

Objetivo

Definir o que é **ciência de dados** e quais são suas principais **técnicas**, **conceitos** e **ferramentas**. De modo a construirmos uma **estrutura teórica** que ressalte possíveis **frentes de pesquisa** e fundamente futuros **aprofundamentos**.

Metodologia

Efetuamos uma revisão sistemática de revisões:

1. **Buscamos** artigos do tipo "Review" contendo o termo "data science".
2. **Analizamos** os títulos e abstracts dos 100 primeiros resultados para **selecionarmos** aquelas que tinham por **objetivo definir ciência de dados**.
3. Finalmente efetuamos uma **análise comparativa** dos 4 artigos selecionados para responder nossas perguntas de pesquisa.



Resultados

Greater Data Science (GDS)

1.

1. Coleta,
2. Preparação
3. Exploração

2.

1. Representação
2. Transformação

3.

1. Computação
de dados

4.

1. Modelagem
de dados

5.

1. Visualização e
apresentação

6.

1. Ciência sobre
Ciência de
Dados

Data Science Work Model (DSMW)

PREPARAÇÃO

1. Definição de
necessidades
2. Coleta de dados
3. Perfilamento
4. Manipulação de
dados

ANÁLISE

1. Experimentação
2. Exploração
3. Modelagem
4. Verificação
5. Interpretação

APLICAÇÃO

1. Monitoramento
2. Refinamento

COMUNICAÇÃO

1. Disseminação
2. Documentação

Discussão

Concluímos que a **Ciência de Dados** se cristalizou a partir dos scripts computacionais de softwares numéricos como a linguagem R. A partir deles, **rotinas** de **coleta**, **processamento** e **análise** de dados puderam ser **compartilhadas**, **reproduzidas** e **escrutinadas** com inédita precisão. Assim, Ciência de Dados é o **domínio** que se ocupa dos **problemas derivados** de todo esse **processo** fundamentalmente interdisciplinar, desde a **definição** de um problema respondível até a **implementação** de uma solução computacional, **interpretação** e **comunicação** dos resultados. Nesse sentido, **DSWM** e **GDS** são modelos representativos e estruturantes da **prática e teoria** da Ciência de dados.

Conclusões

A partir dessa elucidação, é possível direcionar os **próximos aprofundamentos** ao evidenciar as **frentes de pesquisa** em potencial que cada uma dessas sub-áreas representam. Além de **orientar a abordagem prática** de um problema de dados.

Assim, a próxima etapa dessa pesquisa diz respeito justamente ao aprofundamento na área de **Preparação** de dados por meio de um **projeto prático**.