Relatório 6 – O que é Ciência de Dados

Edryck Freitas Nascimento

1. Introdução

O objetivo desta aula foi aprender sobre o que é Ciência de Dados. A tarefa consistia em assistir os vídeos “O que é Ciência de Dados | Nerdologia Tech”[1] e “What REALLY is Data Science? Told by a Data Scientist”[2] e fazer um relatório sobre os conhecimentos adquiridos.

2. Desenvolvimento

Ao assistir os vídeos, pesquisei sobre o artigo “From Data Mining to Knowledge Discovery in Databases” mencionado no vídeo “What REALLY is Data Science? Told by a Data Scientist”, o artigo define as bases do campo de KDD (Knowledge Discovery in Databases), que é a descoberta de conhecimento em banco de dados (a tradução de KDD), onde diferencia da mineração de dados (Data Mining), onde o autor define a mineração de dados como uma das etapas técnicas dentro do processo mais amplo de KDD.

William S. Cleveland, estatístico e cientista da computação, quis trazer a mineração de dados para outro nível combinando Ciência da Computação e mineração de dados, assim resultou em uma estatística muito mais técnica feita por ele, onde ele acreditou que iria expandir as possibilidades de mineração de dados, produzindo uma fonte de inovação. Agora podendo ter a vantagem do poder de um computador para estatística. Ele chama essa combinação de mineração de dados com Ciência da Computação de Ciência de Dados (Data Science).

Próximo desse mesmo período foi quando a internet evoluiu para um modelo colaborativo e interativo, ou seja, substituiu a estrutura estática da Web 1.0 por web como plataforma (chamamos essa evolução de Web 2.0), que permitiu os usuários criar, compartilhar e editar informações. Isso resultou em muitos dados, tantos dados que ficou difícil lidar com tecnologias tradicionais, chamamos isso de Big Data. Isso abriu um mundo de possibilidades em descobrir insights usando dados, mas também significou que a questão mais simples requer uma estrutura de dados mais refinada para suportar o tráfego de dados.

Com a ascensão de Big Data, por volta de 2010 deu impulso a ascensão da Ciência de Dados, ajudando nas demandas dos negócios para traçar insights do massivo e não estruturado conjunto de dados. O *Journal of Data Science* define Ciência de

[1] https://www.youtube.com/watch?v=ykSILAQQu6o
[2] https://www.youtube.com/watch?v=xC-c7E5PK0Y

Dados como (de acordo com o vídeo What REALLY is Data Science? Told by a Data Scientist, não achei a fonte original):

> "[...] quase tudo que tem a ver com dados: coleta, análise, modelagem..., mas a parte mais importante são suas aplicações — todos os tipos de aplicações"

A abundância de dados por volta de 2010, possibilitou treinar máquinas utilizando a abordagem orientada a dados (Data Driven) em vez da abordagem orientada a conhecimento (Knowledge Driven), todos os artigos teórico sobre redes neurais recorrentes e máquinas de vetores de suporte se tornou viável, o aprendizado profundo deixou de ser um conceito acadêmico e se tornou uma classe tangível e útil de aprendizado de máquina, assim o aprendizado de máquina e inteligência artificial dominou a mídia, escondendo os outros aspectos da Ciência de Dados como a análise exploratória, experimentação e habilidade que normalmente são chamadas como inteligência de negócio, fazendo o público geral pensa na Ciência de Dados como pesquisadores focado em aprendizado de máquina e inteligência artificial.

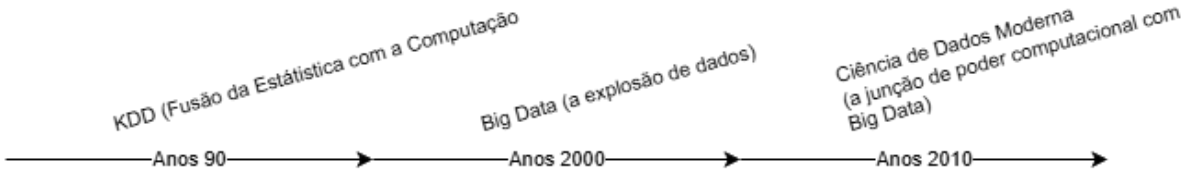

Figura 1. Linha do tempo do surgimento da Ciência de Dados (autoria própria).

A "Hierarquia de Necessidades da Ciência de Dados" mostra bem essa estrutura:

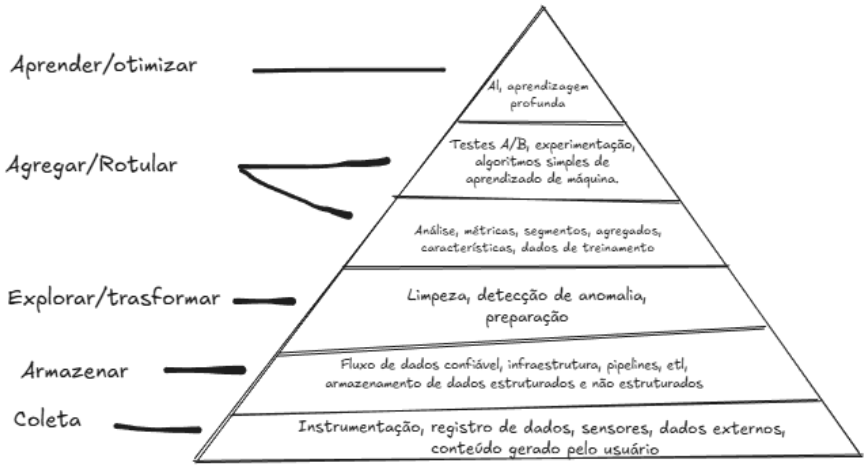

Figura 2. Pirâmide da hierarquia de necessidades da Ciência de Dados. (Robert Chang[3]).

[3] OBS.: Eu não havia encontrado realmente a figura, eu só a redesenhei, o motivo do autor ser Robert Chang é porque era o autor que estava dizendo no vídeo 2.

Na base da pirâmide tem a etapa coleta, pois é preciso coletar algum tipo de dado para utilizar, em seguida mover e armazenar os dados, depois explorar/transformar os dados, essas etapas são conhecidas pela mídia por causa do Big Data. O material intermediário, também o mais importante para as empresas por ser assim que vai estar falando para ela o que vai fazer com o produto, isto é, envolve as análises que informam utilizando os dados, então métricas que vai mostrar o que está acontecendo com o produto, elas dizem se tem sucesso ou não, também testes A/B e experimentação que permite que saiba quais versões do produto são as melhores, essas são coisas muito importante, mas não são abordadas na mídia, o que é abordado na mídia é a parte de aprendizado profundo e inteligência artificial, mas quando pensa sobre isso para uma empresa ou setor, na verdade não vai ser a prioridade mais alta, normalmente não é a coisa que vai produzir mais resultados com o mínimo de esforço, por esse motivo estão no topo da hierarquia de necessidades

A atuação do Cientista de Dados varia conforme o tamanho da empresa:

- **Startups:** normalmente não tem muitos recursos, logo, normalmente vai ter apenas um cientista de dados e ele vai fazer todas as etapas da pirâmide, com exceção do topo dela, porque não é a prioridade no momento, por ter que configurar toda a infraestrutura de dados, tendo que fazer as análises, criar as métricas e fazer os testes A/B sozinho, portanto, para uma startup toda a pirâmide é Ciência de Dados.
- **Empresas de médio porte:** Elas tem muito mais recursos e podem separar os engenheiros de dados dos cientistas de dados, então normalmente a coleta é responsabilidade da engenharia de software, o armazenamento e exploração/transformação é a parte dos engenheiros de dados e por último, o resto da pirâmide tem os cientistas de dados como responsáveis, tendo o cientista de dados mais técnico e assim essas empresas normalmente procuram que tem doutorado ou mestrado na área (normalmente no Brasil não é necessário).
- **Empresas de grande porte:** Elas tem muito mais dinheiro e assim podem gastar mais com funcionários, assim pode ter muitos funcionários diferentes trabalhando com coisas diferentes, dessa forma não precisa pensar nas coisas que não quer fazer e pode se concentrar nas coisas em que ele é melhor, tendo a seguinte divisão da pirâmide de necessidades: engenheiros de software tem cuida da parte de coleta, na exploração e armazenamento tem como responsável os engenheiros de dados, o cientista de dados e analistas ficam responsáveis pela agregação e rotulação de dados e o topo, inteligência artificial e aprendizado profundo, ficam os cientistas de pesquisa, também chamamos de Núcleo de Ciência de Dados, e eles são apoiados pelos engenheiros de ML (aprendizado de máquina/machine learning).

Porte da empresa	Foco do cientista de dados	Principal responsável pela coleta e armazenamento
Startup	Abrange todas as áreas (com exceção do topo)	Cientista de Dados (faz tudo sozinho)
Média	Análise, métricas e testes A/B	Engenheiro de software e dados
Grande	IA, *Deep Learning* e Pesquisas (só o topo da pirâmide)	As equipes são especializadas

Tabela 1. Divisão de responsabilidades de acordo com o tamanho da empresa (autoria própria).

3. Conclusão

A Ciência de Dados não é uma disciplina estática e sim um campo multidisciplinar que evoluiu da estatística e mineração de dados para juntar o Big Data e a Inteligência Artificial. A definição prática da profissão é flexível, dependendo da maturidade de dados da empresa, a pirâmide de necessidades, e dos recursos disponíveis. Sendo assim, um cientista de dados deve entender que seu papel pode variar desde a construção de infraestrutura básica até a criação de modelos mais complexos de IA, isso, claro, dependendo do contexto organizacional onde está inserido.

Referências:

Artigo ***From Data Mining to Knowledge Discovery in Databases***:
https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230

Vídeo ***What REALLY is Data Science? Told by a Data Scientist***:
https://www.youtube.com/watch?v=xC-c7E5PK0Y

Vídeo ***O que é Ciência de Dados | Nerdologia Tech***:
https://www.youtube.com/watch?v=ykSILAQQu6o