FUNDAMENTOS E TÉCNICAS EM CIÊNCIAS DE DADOS

PROF. JOSENALDE OLIVEIRA

josenalde.oliveira@ufrn.br
https://github.com/josenalde/datascience

ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - UFRN

ABRANGÊNCIA DA CIÊNCIA DE DADOS

- Objetivo: melhorar o processo de tomada de decisão, por características/informações (não obvias e úteis) extraídas de grandes conjuntos de dados;

- Percebemos esta realidade em vários segmentos, por exemplo quando:

Recebe-se um anúncio personalizado

Recomendação de filmes, livros e 'amigos'

Filtragem automática (spam list)

Custo de seguros e perfil de clientes

Sequenciamento de semáforos

Chatbots

- Redes sociais
- Poder computacional (custo por MB e por Mbps)
- Métodos eficazes de análise e modelagem, IA
- Acesso à estruturas de coleta, armazenamento e processamento

ABRANGÊNCIA DA CIÊNCIA DE DADOS

MACHINE LEARNING (Aprendizagem de máquina, ML)

Projeto e avaliação de algoritmos para extração de padrões a partir de dados

DATA MINING(Mineração de dados)

Análise de dados estruturados, usualmente com ênfase comercial

Se os dados são textos – mineração de textos, processamento de linguagem natural

A ciência de dados inclui os itens acima, mais o ciclo de dado como um todo, desde a coleta, limpeza e transformação de dados não estruturados (social por exemplo), web tecnologias distribuídas de Big Data e questões éticas de regulação e uso dos dados

Exemplos:

1) Identificar grupos de consumidores com gostos similares (segmentação de mercado): em ciência de dados = CLUSTERING
2) Identificar produtos que são comprados juntos (REGRAS DE ASSOCIAÇÃO)
3) Detecção de eventos anormais, estranhos, como fraudes etc. = DETECÇÃO DE ANOMALIAS (outliers)
4) Classificação (determinar se um dado pertence à determinada CLASSE)
5) Predição simples, com **um único atributo**: se um e-mail contiver no conteúdo a frase "ganhe dinheiro fácil" é SPAM. É algo simples, não é necessário ciência de dados...

ABRANGÊNCIA DA CIÊNCIA DE DADOS

Exemplo inicial de exploração básica de dados textuais – nuvem de palavras

- Métrica: frequência, quantidade de buscas etc.
- Fontes: páginas web, bases de dados diversas
- Google Ads (cliques) num determinado site, Google Console Engine, Answer the Public

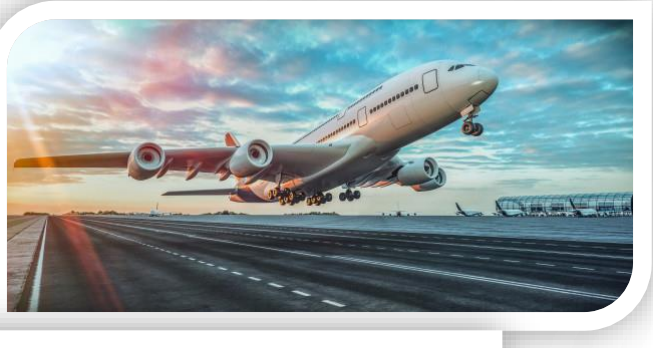
Ver nb3.ipynb

Exemplo de fontes: kaggle, dados.gov.br, dados.ufrn.br (E api.ufrn.br)

CONCEITOS

"Se um especialista humano pode facilmente criar um padrão em sua própria mente, geralmente não vale a pena o tempo e esforço de usar ciência de dados para descobrir tal padrão" (KELLEHER; TIERNEY, 2018)

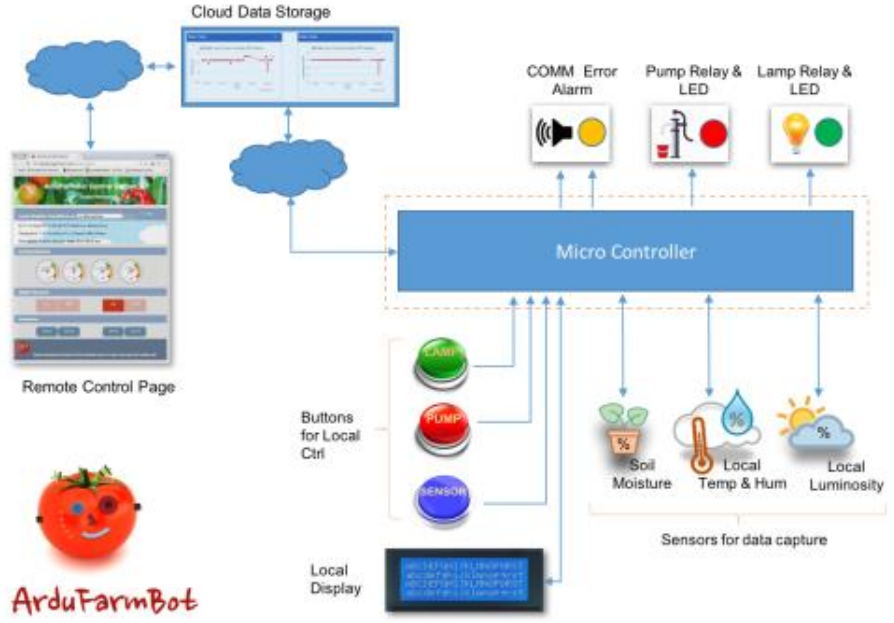
Normalmente se aplica a grandes volumes de dados (datasets) cujos padrões sejam difíceis de extrair manualmente. Normalmente até 3 atributos criamos regras facilmente, após isto pode ser tedioso.



CONCEITOS

Por exemplo um Boeing 787 produz 0,5TB de dados durante um voo, a partir de sensores espalhados pela aeronave. Os sensores nos flaps (extensões da asa para sustentação, usadas no pouso por exemplo) emite sinais de vibração (dado). Embora existam dados não eletrônicos (impressos), o foco está em dados eletrônicos, disponíveis em formato digital.

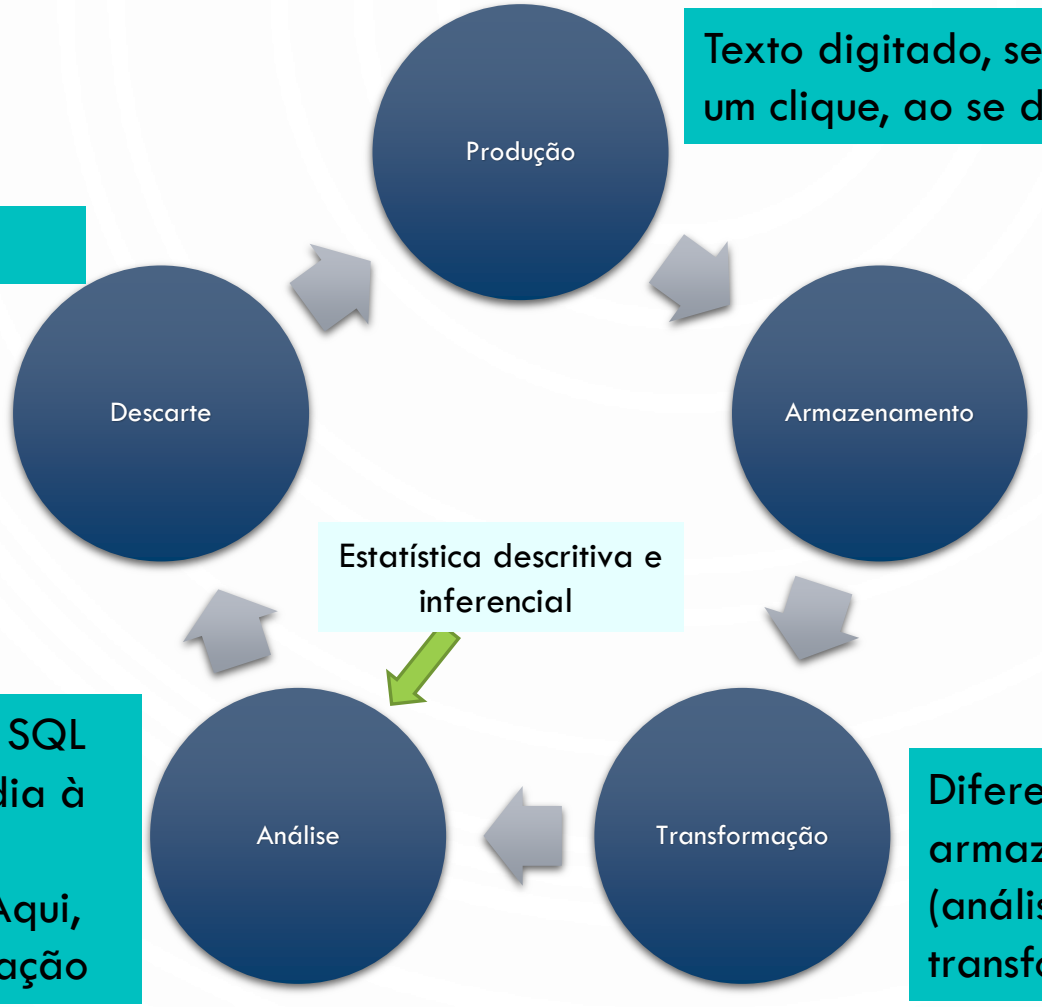
Um dado analógico convertido para digital e persistido torna-se fonte de dados de determinado processo não apenas para monitoramento, mas para 'descoberta' de padrões, identificação de relações



Exemplo de fonte de datasets: Kaggle
https://www.kaggle.com/zynicide/wine-reviews/data?select=winemag-data-130k-v2.csv

CICLO DE VIDA DO DADO

Podemos entender, portanto, ciência de dados como os processos, modelos e tecnologias que estudam os dados durante todo o seu ciclo de vida, da produção ao descarte



Texto digitado, sensor, ao fazer uma foto, um clique, ao se deslocar, gps, ...

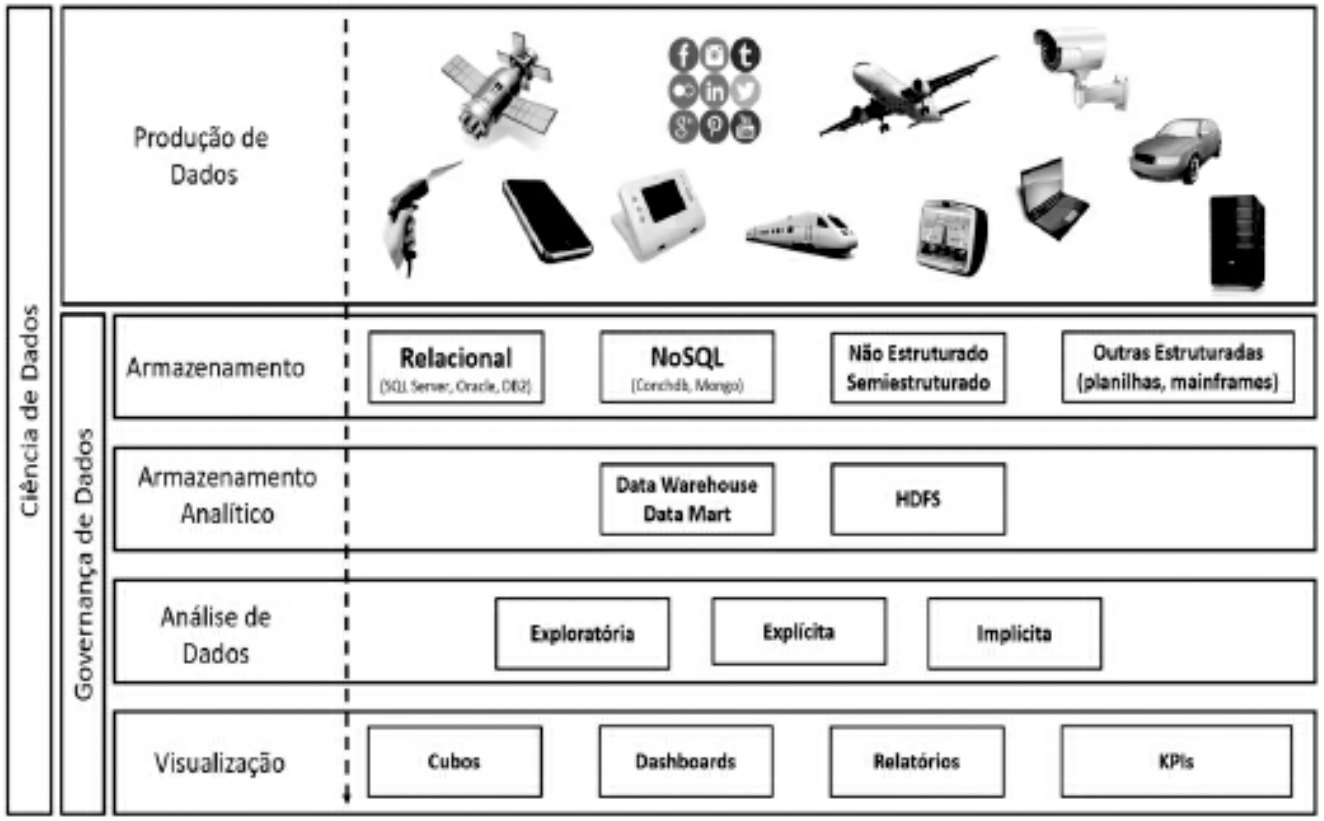
Por quanto tempo manter?

Persistência em alguma estrutura (xml, json, texto plano, bd relacional, não relacional, ...) – com questões de segurança, privacidade, qualidade (policy) – ver projeto PERKEEP: Perkeep

De uma simples consulta SQL para ver as vendas do dia à criação de modelos de classificação com RNA. Aqui, entra também a visualização

Diferença entre modelo de armazenamento e modelo para consumo (análise): ETL em data warehouses, XML transformado nota fiscal 'amigável'

PANORAMA DE CIÊNCIA DE DADOS



Fonte: AMARAL, F. (2018)

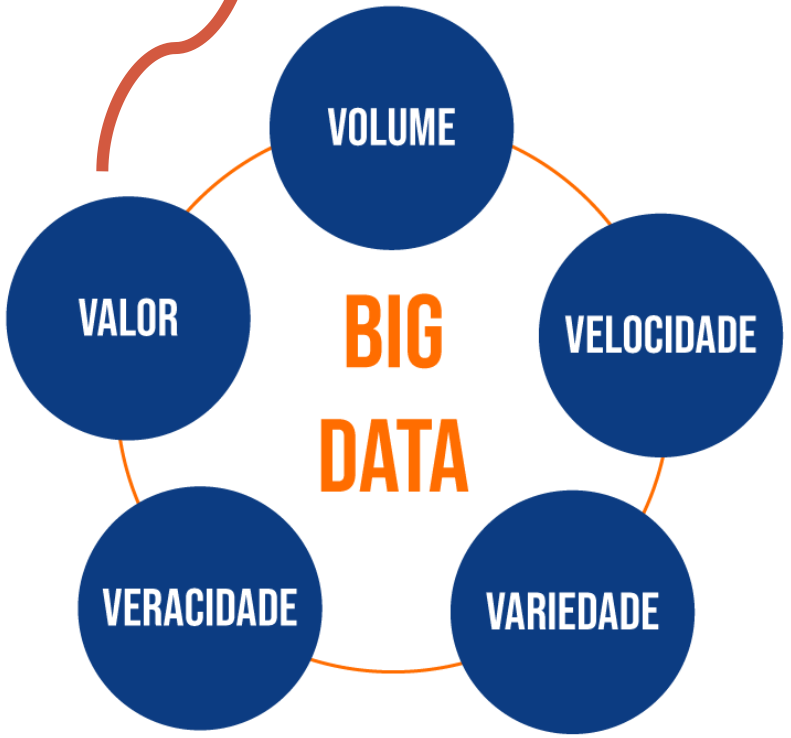
Dados, um ativo que vale ouro!

E O BIG DATA?

DEFINIÇÕES: DE 3 A 5 V'S...

Virtualização
Computação em nuvem (IaaS, PaaS, SaaS)
Sistemas Distribuídos

Para os negócios, vantagem competitiva, Data-driven services, eficiência, antecipação
Minimizar "churn"



Dados são gerados de todas as formas, por todos os lados; Big Data não é uma tecnologia, mas um fenômeno; impacta no conceito de amostragem

A questão é que os eventos sempre existiram e ocorreram, mas os dados não eram REGISTRADOS (datificação)

Um celular hoje com mais capacidade que o melhor PC de 1985
+ de bi de pessoas com telefones
+ de 1,7 bi em redes sociais
+ 3 mi de e-mails por segundo
+ 100 horas de vídeos no YouTube por minuto
+ 500 mi de tweets por dia
+ 100 TB carregados no Facebook etc.