Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

Ciências de Dados e Big Data

Disciplina: Ciências de Dados e Big Data

Horário:

Quarta-Feira – 19:00 – 22:40

Professor :Eduardo Savino Gomes

Email:eduardo.Savino@fecap.br

Ciências de Dados e Big Data

EMENTA: Estudo das atividades que tem por objetivo a coleta, armazenagem, organização, administração, governança e entrega de grandes volumes de dados, tendo em vista assegurar um alto nível de qualidade e **torná-los acessíveis para as aplicações nas áreas de inteligência de negócios e análise de dados**.

Bibliografia

AMARAL, F. Introdução à Ciência de Dados: mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016.

PROVOST, F.; FAWCETT, T. Data Science para negócios. Rio de Janeiro: Alta Books, 2016.

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. Introdução à mineração de dados: com aplicações em R. Rio de Janeiro: Elsevier, 2016.

Superior de Tecnologia em Análise e Desenvolvimento de Sistemas

Ciências de Dados e Big Data

Cronograma Proposto

Aula	Sem		Descrição
1	1	09/ago	Apresentação da Disciplina, Plano de Ensino, Bibliografia, Avaliações
2		09/ago	O que é BigData? O que faz um Cientista de Dados. Ambiente Colab
3	2	16/ago	SISTEMA DISTRIBUÍDO SISTEMA DE ARQUIVOS DISTRIBUÍDOS
4		16/ago	SISTEMA DISTRIBUÍDO SISTEMA DE ARQUIVOS DISTRIBUÍDOS
5	3	23/ago	Hadoop e Spark
6		23/ago	Hadoop e Spark
7	4	30/ago	Introdução ao Python Lista, Matriz
8		30/ago	Introdução ao Python Lista, Matriz
9	5	06/set	Mamipulando dados com Pandas
10		06/set	Mamipulando dados com Pandas
11	6	13/set	Manipulando dados com Pandas e Matplotlib (Estudo de Caso)
12		13/set	Manipulando dados com Pandas e Matplotlib (Estudo de Caso)
13	7	20/set	API – Application Program Interface JSON - JavaScript Object Notation Acessando URL em Python Consumindo API em Python
14		20/set	API – Application Program Interface JSON - JavaScript Object Notation Acessando URL em Python Consumindo API em Python
15	8	27/set	Definindo o Trabalho para NI
16		27/set	Definindo o Trabalho para NI
17	9	04/out	Gráfico de Dispersão Mapa de Calor
18		04/out	Desenvovendo Trabalho NI
19	10	11/out	Desenvovendo Trabalho NI
20		11/out	Desenvovendo Trabalho NI
21	11	18/out	Apresetação do Trabalho NI
22		18/out	Apresetação do Trabalho NI
23	12	25/out	HADOOP MAP REDUCE
24		25/out	HADOOP MAP REDUCE
25	13	01/nov	HADOOP MAP REDUCE SPARK
26		01/nov	HADOOP MAP REDUCE SPARK
27	14	08/nov	HADOOP MAP REDUCE SPARK
28		08/nov	HADOOP MAP REDUCE SPARK
29	15	15/nov	FERIADO - DIA DA PROCLAMAÇÃO DA REPÚBLICA
30		15/nov	FERIADO - DIA DA PROCLAMAÇÃO DA REPÚBLICA
31	16	22/nov	Revisão para PO
32		22/nov	Revisão para PO
25	17	29/nov	PO - PROVA OFICIAL
26		29/nov	PO - PROVA OFICIAL
27	18	06/dez	VISTA DE PROVA
28		06/dez	VISTA DE PROVA
29	19	13/dez	VISTA DE PROVA
30		13/dez	VISTA DE PROVA
31	20	20/dez	EXAME
32		20/dez	EXAME

AVALIAÇÃO:

Atividades teórico-práticas (NI – 20%) e uma prova oficial (PO – 50%).

Atividades de apoio ao Projeto Interdisciplinar (PI – 30%).

Critério de Aprovação: Média > = 6,0 e Frequência Mínima 75%.

NI – 18 DE OUTUBRO

PO – 29 DE NOVEMBRO

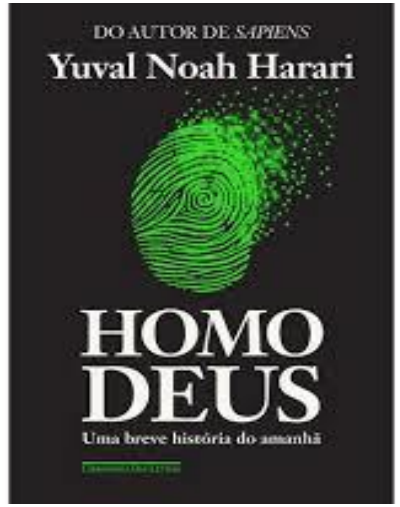
EXAME – 20 DE DEZEMBRO

Ciências de Dados e Big Data

Mas o que é Ciências dos Dados?

A humanidade nunca registrou tantos dados em sua história. *"Estima-se que 90% dos dados armazenados no mundo foram produzidos apenas nos últimos dois anos e os rastros desses dados continuam duplicando a cada ano. "* (Data Science Academy.)

De acordo com Yuval Harari vivemos na era do **dataísmo**, ou seja, uma espécie de religião dos dados



Ciências de Dados e Big Data

Mas o que é Ciências dos Dados?

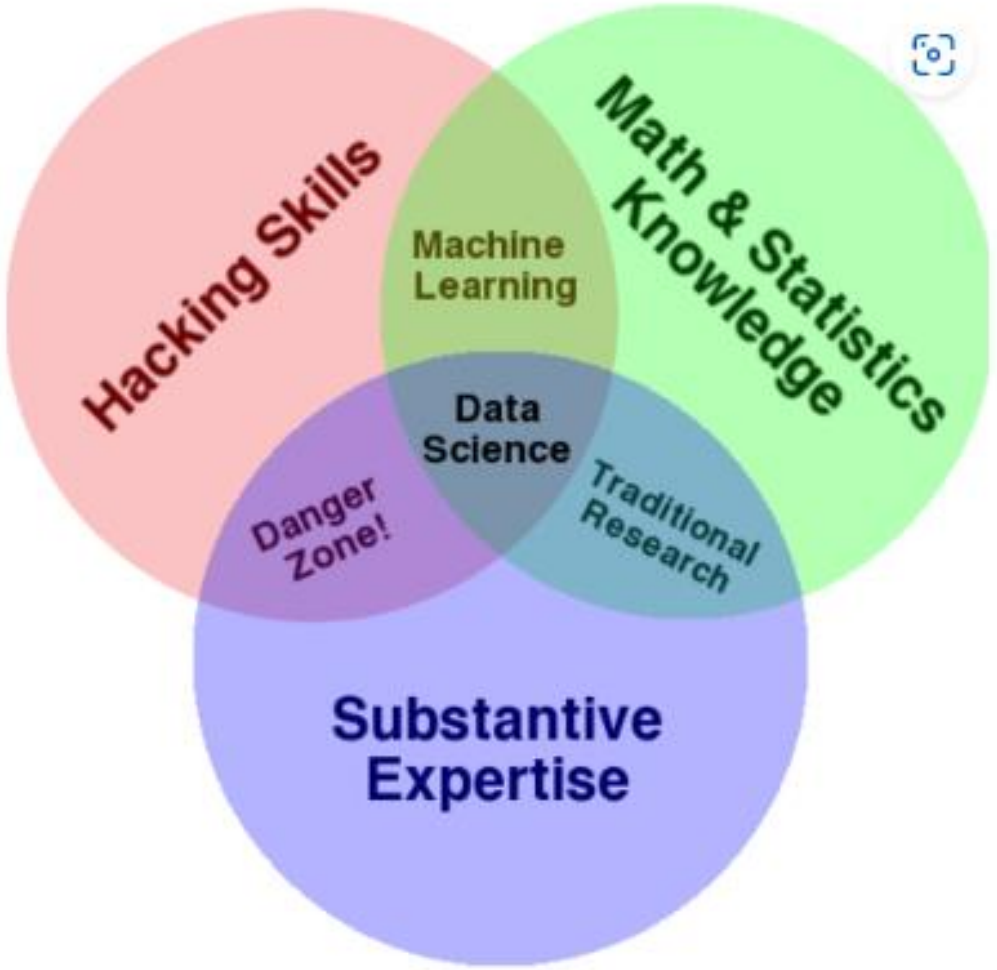
Não ´há uma definição de consenso sobre o que é Ciência de Dados, afinal toda ciência é baseada em dados (Certo?). Porém uma definição que podemos utilizar é:

"Ciência de Dados é o processo para extrair informações valiosas a partir de dados. Como estamos vivendo na era do Big Data, a Ciência de dados está se tornando um campo muito promissor para explorar e processar grandes volumes de dados gerados a partir de várias fontes e em diferentes velocidades." (Data Science Academy.)

Ciências de Dados e Big Data

Mas o que é Ciências dos Dados?

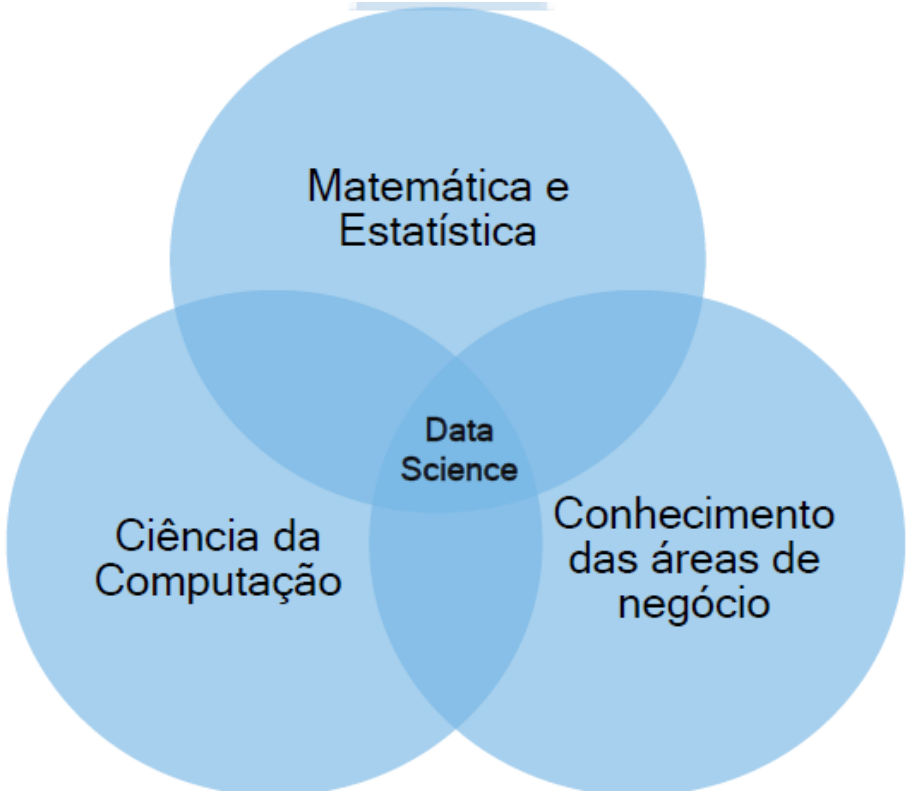
Apesar de não haver um consenso sobre a definição de Ciências dos Dados talvez este seja o melhor termo que temos para expressar um atividade que exige conjunto interdisciplinar de habilidades. Dessa forma um boa definição existente de ciência de dados é ilustrada pelo *Data Science Venn Diagram de Drew Conway*, publicado pela primeira vez em seu blog (Lab — Drew Conway).



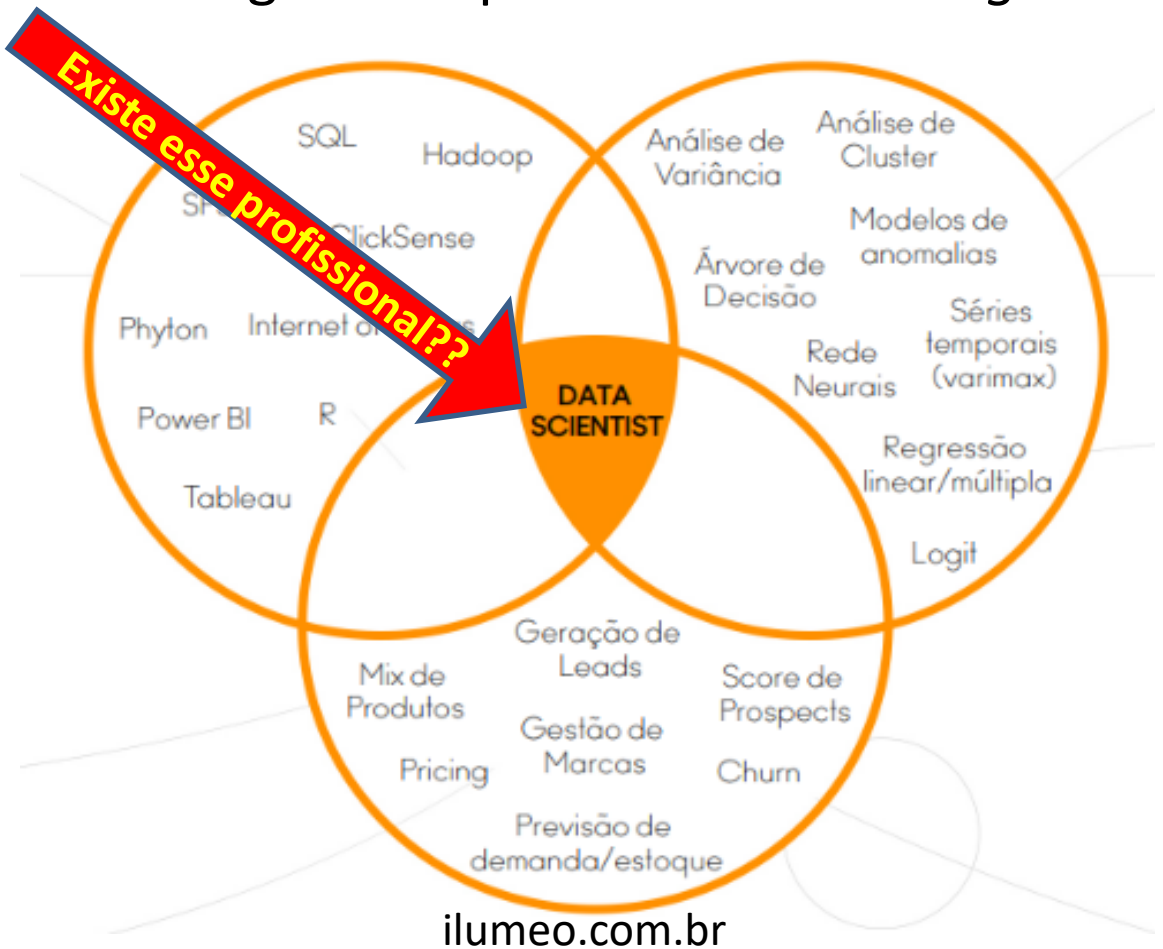
Ciências de Dados e Big Data

Mas o que é Ciências dos Dados?

De maneira mais específica outros diagramas surgiram a partir do *Venn Diagram de Drew Conway*



Data Science Academy



ilumeo.com.br

Ciências de Dados e Big Data

Então o que faz um Cientista de Dados?

Há uma piada (*sem graça*) que diz que um Cientista de Dados é alguém que sabe mais sobr estatística do que um cientista da computação e mais sobre ciência da computação do que ur estatístico.

Um cientista de dados utiliza de métodos automatizados (ciência da computação) para analisar enormes quantidades de dados (estatística) para extrair conhecimento (áreas de negócio) a partir deles que irão suportar tomadas de decisão.

Ciências de Dados e Big Data

Mas Então o que são Dados?

Dados x Informação x Conhecimento

Dados: Fato, independe do observador para existir.

Ex. Placa de Trânsito.

Informação: Significando ou Interpretando o Dado. Depende do Observador

Conhecimento: Transforma uma Informação em uma ***Ação*** ou Outra ***Informação***

Ciências de Dados e Big Data

Fonte de Dados

Dados Estruturados (Banco de Dados Relacionais)

MySQL, Microsoft SQL Server, IBM DB2, ORACLE

Dados Não Estruturados

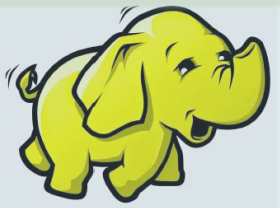
Redes Sociais

Outras fontes e Tecnologias

NoSQL Not Only SQL

CouchDB, Redis, Neo4j, Cassandra, mongoDB, membase, riak

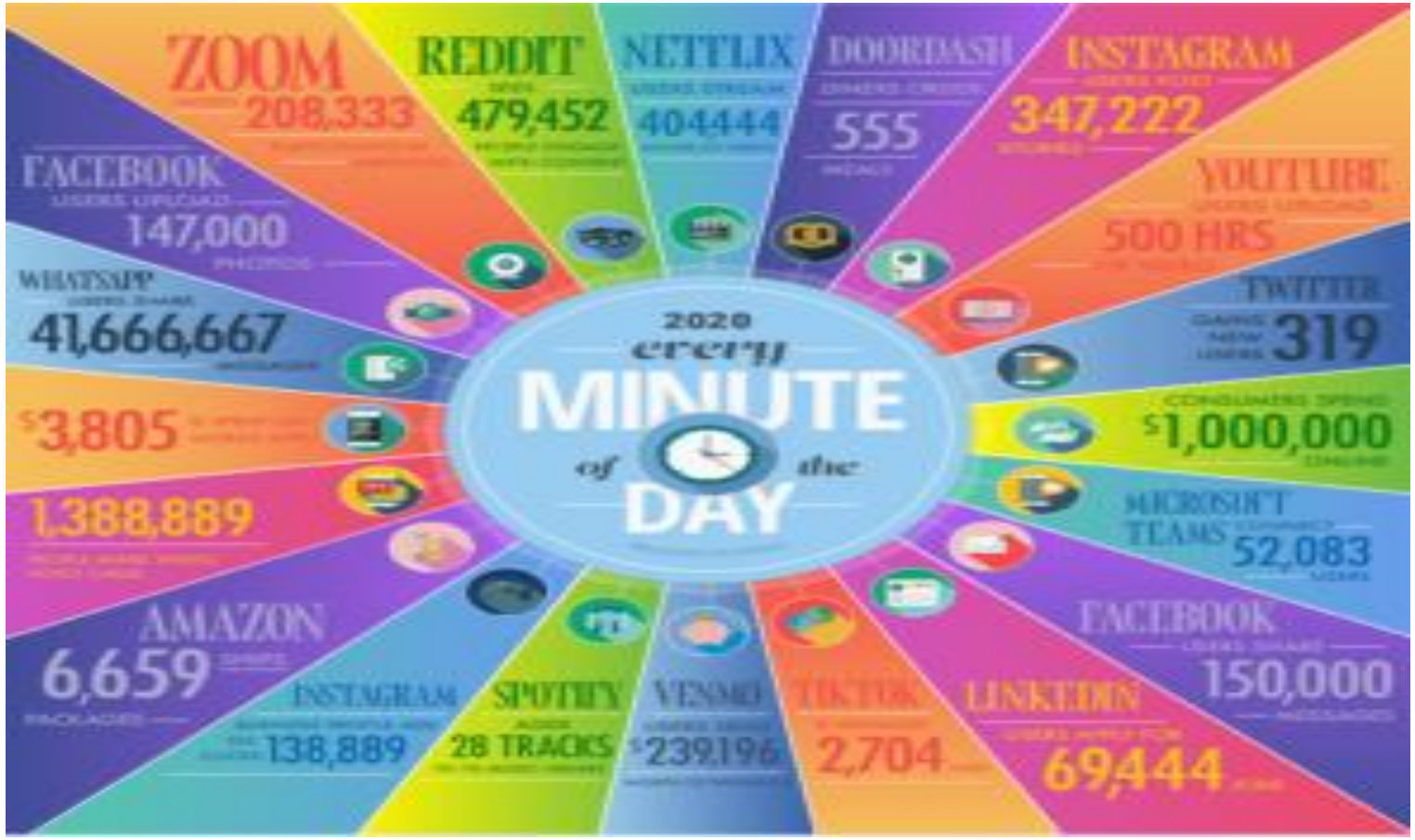
Hadoop



cloudera Ask Bigger Questions, Hortonworks, MAPR TECHNOLOGIES

Ciências de Dados e Big Data

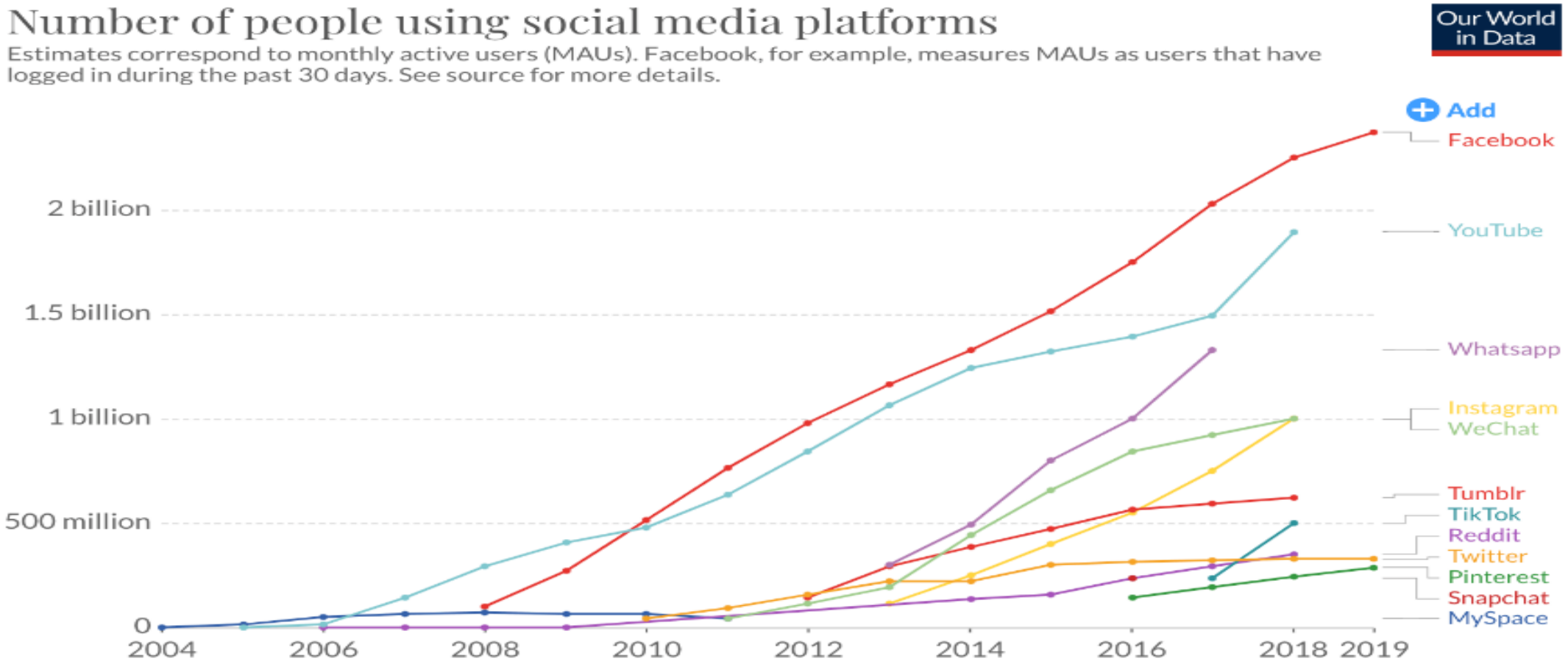
Mas Então o que é BIG DATA?



(https://www.datanami.com/2020/09/04/10-big-data-statistics-that-will-blow-your-mind/)

Ciências de Dados e Big Data

Mas Então o que é BIG DATA?



(https://www.smartinsights.com/social-media-marketing/social-media-strategy/new-global-social-media-research/)

Ciências de Dados e Big Data

Big Data

A quantidade e a forma como os dados são gerados e tratados atualmente levou a criação do termo BIG DATA

Big Data

- **Big Data**

Big Data Analytics é a área do conhecimento que estuda como tratar, coletar, analisar e obter "*informações*" a partir de conjuntos de dados grandes demais para serem analisados por sistemas tradicionais.

Volume: relacionado a grande quantidade de dados gerados;
Variedade: as fontes de dados são muito variadas, o que aumenta a complexidade das análises;
Velocidade: Devido ao grande volume e variedade de dados, todo o processamento deve ser ágil para gerar as informações necessárias;
Veracidade: A veracidade está ligada diretamente ao quanto uma informação é verdadeira.
Valor: Este conceito está relacionado com o valor obtido desses dados, ou seja, com a "informação útil".

Big Data

Volume

Atualmente produzimos mais dados por dia do que se produziu em todos os tempo até alguns poucos anos atrás. Assim, torna-se necessário tratar esse grande volume de forma diferenciada do que a forma atual. Bancos de dados relacionais e modelos ROLAP não suportam mais esses grandes volumes de forma satisfatória.

Big Data

Velocidade

Analisar dados históricos não é mais suficiente para alguns tipos de tomadas de decisão. As fraudes ocorrem a todo momento, quanto mais rápido a Sefaz (Secretaria da Fazensa) conseguir identificar as fraudes praticadas por empresas laranja, menor será a perda para a administração pública. Analisar dados em tempo real já é uma realidade.

Big Data

Velocidade

Analisar dados históricos não é mais suficiente para alguns tipos de tomadas de decisão. As fraudes ocorrem a todo momento, quanto mais rápido a Sefaz (Secretaria da Fazensa) conseguir identificar as fraudes praticadas por empresas laranja, menor será a perda para a administração pública. Analisar dados em tempo real já é uma realidade.

Big Data

Variedade

Os tipos de informações a serem analisadas em processos decisórios não se limitam mais somente aos dados históricos vindo de bancos de dados relacionais, é preciso considerar os dados **não estruturados** originários de mídias sociais , emails, pdfs, documentos eletrônicos, planilhas, etc.

Big Data

Veracidade

Dados devem ser autênticos e devem fazer sentido no contexto de sua análise.
No contexto da Big Data, onde as fontes de dados são das mais diversas, garantir a veracidade e seu devido contexto não é algo trivial.

Big Data

Valor

É necessário que a implementação de um projeto dessa natureza retorne o investimento realizado, ou seja, informação tem valor e esse valor deve saltar aos olhos em retorno de um projeto Big Data.

Considerando o grande volume de dados e o investimento para seu armazenamento e processamento, deve-se garantir que de fato os dados armazenados e processados podem gerar valor.

Ciências de Dados e Big Data

Ferramentas para Ciências de Dados

Google colab

NumPy

matplotlib

pandas