BIG DATA

GUTO CHRISTIANINI


O que é Big Data?
Big Data é um termo utilizado para descrever grandes volumes de dados e que ganha cada vez mais relevância à medida que a sociedade se depara com um aumento sem precedentes no número de informações geradas a cada dia.
As dificuldades em armazenar, analisar e utilizar grandes conjuntos de dados têm sido um considerável gargalo para as companhias.

By IBM

BIG DATA

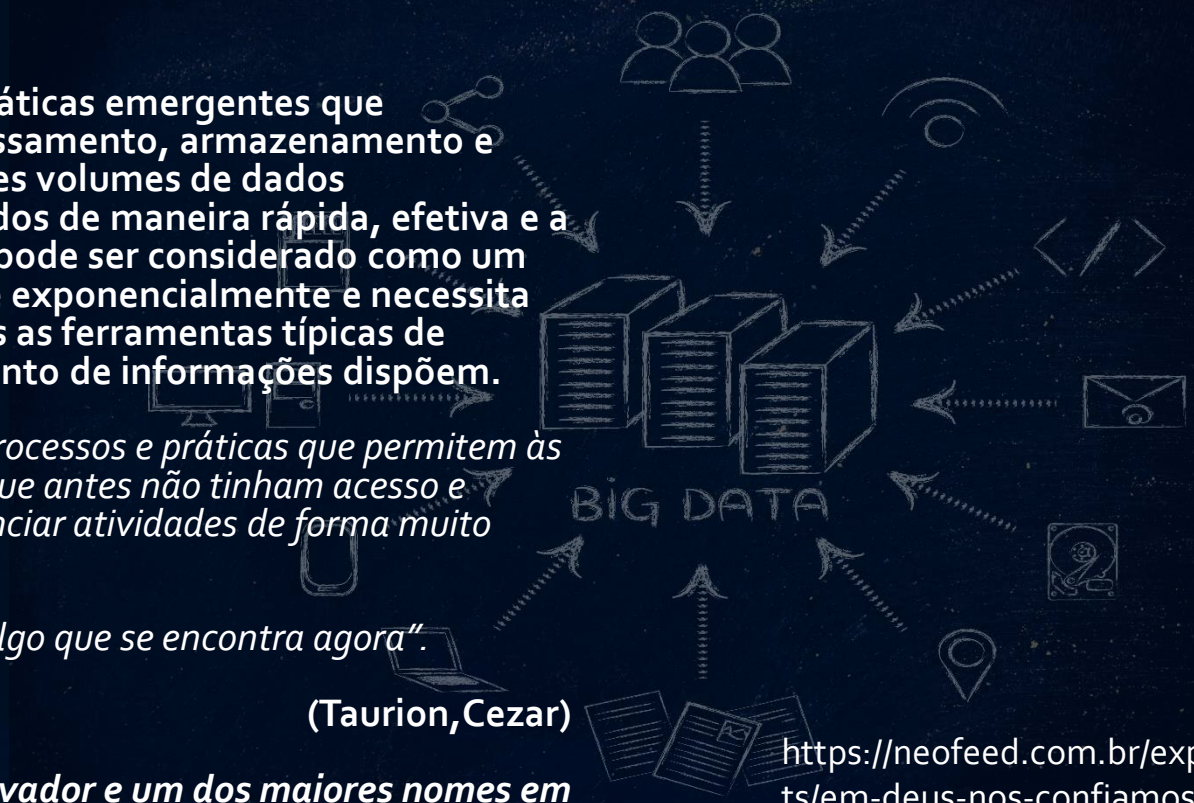
Big Data são tecnologias e práticas emergentes que possibilitam a seleção, processamento, armazenamento e geração de insights de grandes volumes de dados estruturados e não estruturados de maneira rápida, efetiva e a um custo acessível. Big Data pode ser considerado como um conjunto de dados que cresce exponencialmente e necessita de habilidades além das quais as ferramentas típicas de gerenciamento e processamento de informações dispõem.

"Um conjunto de tecnologias, processos e práticas que permitem às empresas analisarem dados a que antes não tinham acesso e tomar decisões ou mesmo gerenciar atividades de forma muito mais eficiente".

"não é teoria ou futurologia, é algo que se encontra agora".

(Taurion,Cezar)

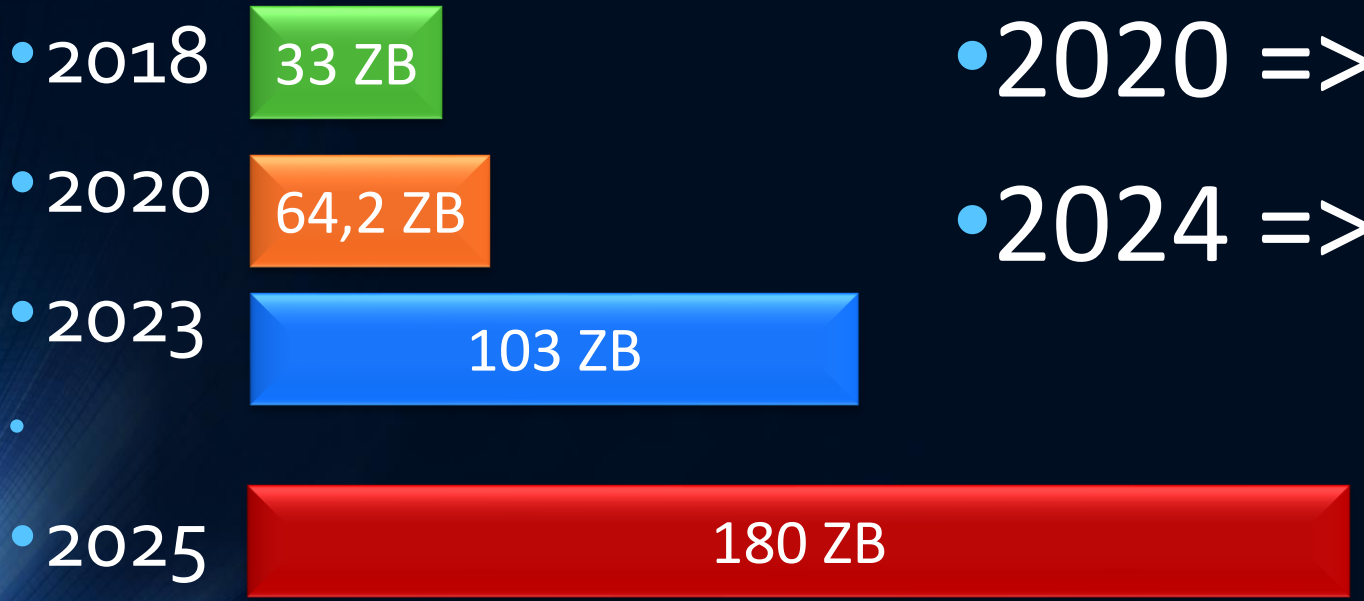
Empreendedor, inovador e um dos maiores nomes em Tecnologia da Informação no Brasil



https://neofeed.com.br/experts/em-deus-nos-confiamos-todos-os-outros-tragam-dados/

VOLUME DE NOVOS DADOS (IDC)

A proporção de dados exclusivos (criados e capturados) para dados replicados (copiados e consumidos)



- 2020 => 1 : 9
- 2024 => 1 : 10

1 ZB = 1 trilhão de Gigabytes

Global DataSphere da International Data Corporation (IDC)

As principais descobertas da previsão Global DataSphere

- A quantidade de dados criados nos próximos três anos será maior do que os dados criados nos últimos 30 anos, e o mundo criará mais de três vezes os dados nos próximos cinco anos do que nos cinco anteriores.
- Dados de produtividade / incorporados são a categoria de criação de dados de crescimento mais rápido, com um CAGR (Tx de crescimento anual composta) de 40,3% para o período de previsão de 2019–2024.
- Em 2024, os dados de entretenimento serão 40% do *Global DataSphere* e a produtividade / dados incorporados serão 29%, paralisados pela dinâmica do COVID-19.
- Sensores estão sendo embutidos em tudo e estão liberando dados que podem ajudar a contextualizar os dados. Esses dados, juntamente com quantidades crescentes de metadados (dados sobre dados), estão crescendo agressivamente e em breve ultrapassarão todos os outros tipos de dados.
- A participação do consumidor no *Global DataSphere* ficará em torno de 50% e diminuirá cerca de 4% nos próximos cinco anos, lentamente cedendo participação ao DataSphere corporativo.

Dados sobre "Dados"

- **Volume de Dados**:
  - As empresas geram cerca de **2 quintilhões de bytes de dados por dia** em todos os setores. Esses dados têm um valor estimado de **US$ 77 bilhões em 2023** e continuam crescendo[1].
  - Em 2025, especialistas indicam que mais de **463 exabytes de dados** serão criados diariamente, o que equivale a aproximadamente **212.765.957 DVDs** [1].
- **Impacto Econômico**:
  - A **baixa qualidade dos dados** pode custar à economia dos EUA até **US$ 3,1 trilhões por ano** [1].

Dados sobre "Dados"

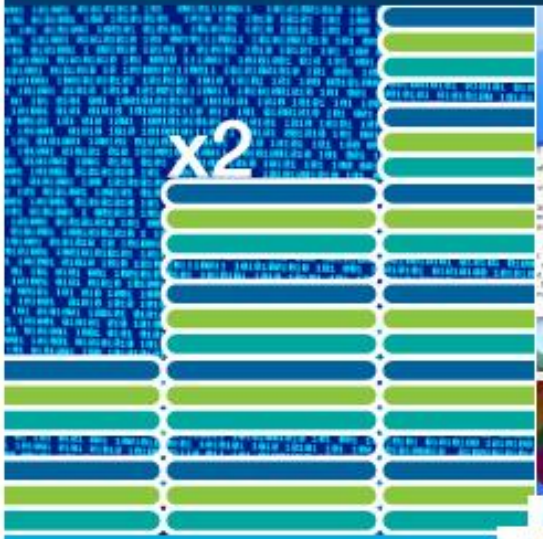
- **Investimento em Big Data**:
  - O mercado de análise de Big Data está projetado para atingir um valor de cerca de **US$ 103 bilhões até 2027** [1].
  - **97,2% das organizações** afirmam que estão investindo em **IA e Big Data** [1].
- **Desafios**:
  - Cerca de **95% das empresas** afirmam que sua incapacidade de entender e gerenciar **dados não estruturados** as está impedindo [1].
  - Apenas cerca de **26% das empresas** afirmam ter alcançado uma **cultura orientada a dados** [1].

"Vivemos em um mundo cada vez mais habilitado e assistido por vídeo e consumimos uma quantidade cada vez maior de vídeos de entretenimento a cada ano - esses são os principais fatores que impulsionam o crescimento do Global DataSphere"

John Rydning , vice-presidente de pesquisa Global DataSphere da IDC

MUITAS OUTRAS....

Big Data



Volume

Geramos um número gigantesco de dados diariamente, e estima-se que esse volume dobre a cada 18 meses.

Variedade

Esses dados vêm de sistemas estruturados e não estruturados gerados por emails, postagens em mídias sociais, mensagens instantâneas, etiquetas RFID, câmeras de vídeo, etc.

Velocidade

Muitas vezes precisamos agir em tempo real para lidar com essa imensa quantidade de dados.

3 Vs


3 Vs DO BIG DATA
1
VOLUME
• TERABYTES;
• RECORDS;
• TRANSAÇÕES;
• ARQUIVOS.
2
VELOCIDADE
• BATCH;
• TEMPO APROXIMADO;
• TEMPO REAL;
• STREAMS.
3
VARIEDADE
• ESTRUTURADA;
• NÃO ESTRUTURADA;
• SEMIESTRUTURADA;
• TODOS ACIMA.

5Vs – Existem autores que consideram+2

- **Volume** – Cada vez mais, produzimos informações em maior quantidade;
- **Variedade** – Com as diversas plataformas e meios de comunicação, as fontes de dados são mais variadas;
- **Velocidade** – Com o avanço das tecnologias, a produção de dados é mais veloz e a tomada de decisão mais rápida torna-se cada vez mais importante.
- **Veracidade** – Garantia de que os dados utilizados estão corretos e são válidos;
- **Valor** – Garantia de que os dados utilizados agreguem valor para o negócio.

5 Vs


OS 5 Vs DO BIG DATA
1 VOLUME
2 VELOCIDADE
3 VARIEDADE
4 VERACIDADE
5 VALOR

O que são dados estruturados e não estruturados?

- O BIG DATA é um grande banco de dados .. OK!
- O BIG DATA possui uma variedade de dados, todos os tipos e estruturas de dados... OK!
- O BIG DATA se atualiza e aumenta em VOLUME em uma VELOCIDADE rápida e constante ou progressiva.. OK!
- O BIG DATA precisa ter dados confiáveis, a VERACIDADE é ponto fundamental, "pior do que não ter a informação, é ter a informação errada" ... OK!
- O BIG DATA deve ser construído com propósito, para entregar VALOR.. OK!

DADOS ESTRUTURADOS

- Dados estruturados são aqueles que contam com uma estrutura determinada e apresentam informações úteis sobre o perfil dos clientes que as empresas querem atingir, como localização, vendas, contatos, entre outros.
- Estão estruturados em Bancos de Dados relacionais e possuem estruturas bem definidas, rígidas, desenhadas antes de se ter o dado inserido, ou seja, estruturas preparadas para guardarem/carregarem dados específicos e com critérios.
- Exemplos:
  - Sistemas financeiros, de recursos humanos, ERP, PDV
  - Ao negar ou aceitar um cartão de crédito para um correntista, o banco não está fazendo nada mais do que buscar o perfil de consumo e pagamento dele com base nas informações disponíveis.

DADOS NÃO ESTRUTURADOS

- Os dados não estruturados são aqueles que chegam sem nenhuma definição, e é necessário catalogar todas as informações recebidas nas mídias sociais, como Facebook, Instagram, YouTube, Twitter, portais de notícia, entre outros.
- Os **dados não-estruturados**, como o próprio nome já diz, **não** possuem estrutura de organização, sendo totalmente desestruturados. E os **dados** semiestruturados, estão posicionados entre os extremos, **não** possuem uma forma rígida, mas também **não** são totalmente sem qualquer estrutura.
- Exemplos:
  - Comentários em redes sociais, postagens, vídeos, podcast, etc.
  - **Dados não estruturados** corresponde a **80%** dos **dados** corporativos, podendo ser encontrados na forma de **e**-mails, comentários em redes sociais, vídeos, entre outros.

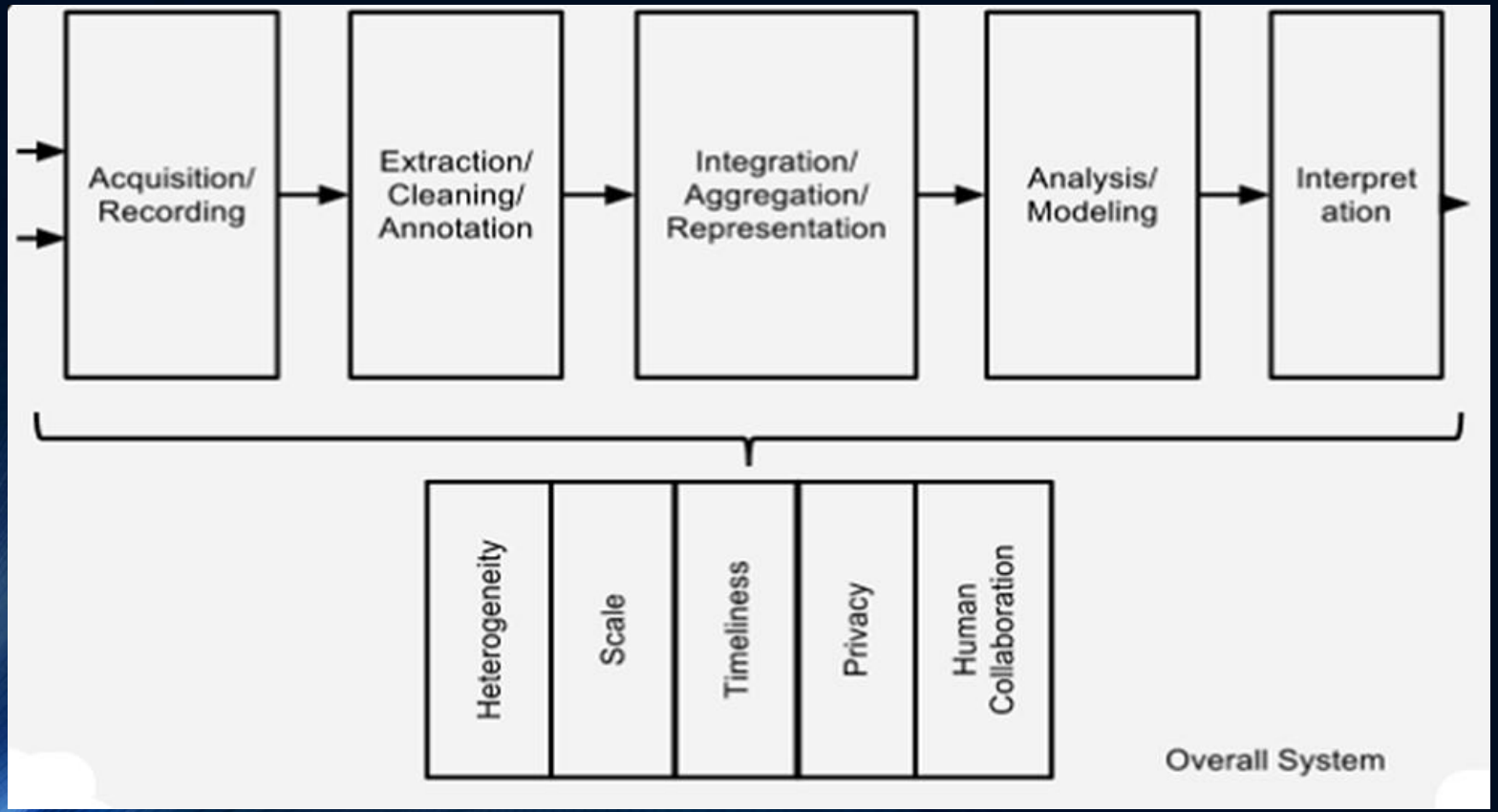
DADOS NÃO ESTRUTURADOS

- Como funciona:
- Primeiro é realizado um monitoramento nos canais de comunicação para extrair comentários sobre determinado assunto ou empresa com base nas palavras-chave. Sendo assim, é possível que o profissional analise tais informações, filtre as ironias, sarcasmo e deboches que a análise de um computador, por exemplo, pode deixar passar despercebido.
- Após a filtragem, toda a informação alcançada deve ser catalogada e separada em tags para facilitar uma futura busca sobre o mesmo assunto. Por causa dessas particularidades, os trabalhos com os dados não estruturados são considerados mais complexos do que com os estruturados.

Os elementos do **Big Data** incluem:

- O grau de complexidade de um conjunto de dados;
- O montante do valor que pode ser obtido utilizando técnicas de análises/contextualização inovadoras vs. não-inovadoras;
- O uso de informação longitudinal dentro da análise e latitudinal na contextualização.
- Tamanho é a primeira definição de **Big Data**. A resposta está no número de fontes independentes de dados e no potencial de interação destes.
- O **Big Data** não se deixa domar pela utilização de técnicas padronizadas de gerenciamento da informação, simplesmente por conta de sua característica de combinações inconsistentes e imprevisíveis.

FASES DO PROCESSO DE BIG DATA - Sugerido



VOCÊ SABIA ?

- ...O furacão Katrina foi previsto –em toda sua violência- com 48 horas de antecedência? A falta de preparo logístico para assimilar a informação garimpada no BIGDATA levou à morte 1600 pessoas.
- ...As agências de risco norte-americanas previram que a chance de ocorrência do estouro da bolha imobiliária em 2008 era de 0,12%? E erraram feio. O risco real era de 28%
- ...No universo BIG DATA as raposas se saem bem melhor que os porcos-espinho?

SEGUNDO A VEJA

- Quatro em cada dez empresas que possuem um departamento de tecnologia já adotam técnicas de **Big Data** em seus negócios e 83% delas confirmam que a novidade ajudou a aumentar os lucros;
- Varejistas que usam táticas de **Big Data** para alavancar negócios veem o lucro crescer, em média, 60%;
- Governos que utilizam a tecnologia economizaram -ao todo- US$330 bilhões em gastos anuais, ao tornar a administração pública mais eficiente;
- O déficit de mão de obra especializada em **Big Data** poderá prejudicar os avanços .

CURIOSIDADES

- Cerca de **90%** dos dados no mundo foram criados nos últimos dois anos.
- O Google processa mais de 40.000 consultas de pesquisa a cada segundo.
- Uma pessoa gera, em média, 1,7 megabytes de dados por segundo durante o uso da internet.
- O Big Data pode ser usado para prever surtos de doenças com precisão, permitindo que as autoridades de saúde pública ajam rapidamente para conter a propagação de uma doença.
- O Facebook armazena cerca de **300 petabytes** de dados gerados por seus usuários todos os dias, o que equivale a cerca de 300 milhões de gigabytes.
- O Google utiliza Big Data para personalizar anúncios de acordo com as preferências dos usuários, tornando os anúncios mais relevantes e aumentando a eficácia do marketing digital.

DESAFIO – Profissionais Qualificados

- A EMC Brasil realizou uma pesquisa onde **73%** das empresas entrevistadas apontaram a **cultura** como sendo a maior barreira de lidar com o Big Data.

- O levantamento destaca que **88%** das companhias acreditam que será um desafio capacitar seus trabalhadores para a nova TI.

"Não está fácil encontrar profissionais de TI. E a dificuldade para Big Data é tamanha porque o conceito vai além dos dados armazenados na TI tradicional".

Carlos Cunha, diretor geral da EMC Brasil

Ferramentas usadas em BIG DATA

- NoSQL
  - DatabasesMongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- MapReduce
  - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum
- Storage
  - S3, Hadoop Distributed File System
- Servers
  - EC2, Google App Engine, Elastic, Beanstalk, Heroku
- Processing
  - R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop

Passo a Passo para Projetos BIG DATA

1. **Definir as necessidades de negócio**

   Definição de necessidades de informações que o negócio precisa melhorar, impulsionar ou sustentar, onde estão os processos críticos e o que é preciso resolver com um projeto de dados

2. **Otimizar a estrutura de dados**

   Integrar e entender as estruturas de dados, imprimindo confiabilidade nas informações e sistemas que as coletam

3. **Defina o que é relevante**

   Revestimento de estruturas e quadros de governança, a fim de permitir a identificação e avaliação eficazes e atentas aos riscos. Tudo atrelado à definição de quais dados são de fato relevantes

Passo a Passo para Projetos BIG DATA

4. **Pense grande e comece pequeno**

   Identificação de fraquezas e fortalezas (negócio, equipe, tecnologia, metodologia, etc). Início de pequenos projetos orientados para a experiência, para aprendizados e melhorias

5. **Estabeleça e monitore a eficiência do seu programa de dados**

   Mensuração de sucesso do projeto e otimização contínua do programa de Big Data, acompanhando o status da integração, riscos, problemas e oportunidades.

   É Importante monitorar o impacto no desempenho das empresas a partir da integração da tecnologia adquirida com as plataformas existentes

VIDEOS E FONTES

- olhar digital - https://www.youtube.com/watch?v=FTlkrLq1pg8
- NETFLIX CPBR8 - https://www.youtube.com/watch?v=b5amQu7q3kA
- Mkt – https://www.youtube.com/watch?v=mPOQTskNEbk
- Nerdologia-https://www.youtube.com/watch?v=hEFFCKxYbKM
- CODFONTE TV - https://www.youtube.com/watch?v=Jqard9dCoWE
- Casas Bahia-https://www.youtube.com/watch?v=OoEdtj90eeE
- CAPPRA - https://youtu.be/AR6EbH7QALg

ARTIGO

- NETFLIX - https://www.bimachine.com.br/post/como-a-netflix-usa-o-big-data
- DEVMEDIA - https://www.devmedia.com.br/big-data-tutorial/30918

Obrigado

PROF. GUTO CHRISTIANINI
JOSE.CHRISTIANINI@FATEC.SP.GOV.BR