# Revisão bibliográfica sobre as tecnologias de machine learning e mineração de dados massivos

Ráfagan Soares de Carvalho Sampaio Mariz
*Departamento de Ciência da Computação*
*Universidade Federal de Minas Gerais*
*Belo Horizonte, Brasil 31270-901*

Vinícius de Oliveira Silva
*Departamento de Ciência da Computação*
*Universidade Federal de Minas Gerais*
*Belo Horizonte, Brasil 31270-901*

***Resumo*—Essa revisão bibliográfica tem o objetivo de compilar artigos, livros e referências online em duas áreas em ascensão dentro da Ciência da Computação: Mineração de dados e aprendizado de máquina. Pretende-se apresentar definições e modelos de cada área, bem como as ferramentas que dão suporte aos algoritmos desenvolvidos para resolver problemas desafiadores em ambos os campos. Ao final, apresentaremos alguns cenários e casos de uso na atualidade que utilizam as técnicas mencionadas. Pretende-se também discutir aspectos éticos e técnicos que permeiam o tema e introduzir o leitor aos problemas resolvidos por essas técnicas e às questões que vêm sendo levantadas a respeito do uso das mesmas nos mais diferentes cenários.**

## 1. Introdução

O crescimento expressivo da geração de dados está cada vez mais perceptível e vem sendo mais e mais discutido em nossa sociedade. O potencial de aplicação desses dados vem atraindo a atenção de diversos setores da economia e, com isso, técnicas têm se mostrado necessárias para que seja possível explorar o poder e a informação que podem ser obtidos de grandes volumes de dados. Os algoritmos de *Machine Learning* (ML) e de mineração de dados são capazes de fornecer *insights* e uma modelagem para vários tipos de bases de dados, abrangendo a indústria, a pesquisa e diversas outras esferas da sociedade.

Inicialmente, algoritmos de aprendizado de máquina foram desenvolvidos e aplicados em ambientes acadêmicos, com grande foco em pesquisa. A recente explosão de dados e o advento de ferramentas poderosas o suficiente para processá-los, porém, fizeram com que o aprendizado de máquina passasse a ser utilizado em uma grande variedade de aplicações, atingindo o setor empresarial e também o consumidor final. Essas novas demandas criaram desafios inéditos para o processamento de dados, e isso levou ao desenvolvimento de muitas pesquisas, especialmente no sentido de desenvolver técnicas escaláveis para tratar um volume de dados sem precedentes na história da computação. Várias ferramentas desenvolvidas, e algumas delas serão discutidas nesta revisão.

Este artigo está dividido da seguinte forma: A seção 2 trata a respeito do aprendizado de máquina de maneira generalista, a seção 3 aborda temas relacionados a processamento de dados massivos e mineração de dados, a seção 4 se encarrega de mostrar como essas duas áreas de conhecimento podem se complementar, detalhando seus algoritmos e algumas ferramentas selecionadas (*TensorFLow*, *Scikit-learn*, Mahout, MLIB e SAMOA); na seção 5, casos de uso e exemplos de aplicações são abordados; e a seção 6 dedica-se a fornecer uma conclusão reflexiva a respeito das contribuições apresentadas.

Processamento de Dados Massivo em Nuvem
11 de dezembro, 2017

## 2. *Machine Learning* e processamento de dados massivos

O objetivo do aprendizado da máquina é permitir que um sistema aprenda do passado ou presente e use esse conhecimento para fazer previsões ou decisões sobre eventos futuros desconhecidos. Em termos gerais, o fluxo de trabalho para uma tarefa de aprendizagem de máquinas consiste em três fases: construir o modelo, avaliar e ajustar o modelo e, em seguida, colocar o modelo em produção. O núcleo da aprendizagem de máquina utiliza os dados que alimentam os modelos e a nova era da Big Data alavanca a aprendizagem de máquinas, colocando ela na vanguarda das pesquisas e aplicações industriais. O significado do termo ”Big Data” ainda é um assunto que gera algum desentendimento e debates, mas geralmente se refere a base dados muito grandes ou muito complexas para processar em apenas uma única máquina, requerendo assim que se execute os algoritmos em diversas máquinas em paralelo. As explicações mais amplamente aceitas tendem a descrevê-lo em termos dos desafios que ela apresenta. Isso é às vezes referido como o ”o problema do Big Data”. Em 2001, Doug Laney [1] descreveu três dimensões de desafios de gerenciamento de dados. A caracterização dos três V's na figura 1 [2] (volume, velocidade e variedade) é frequentemente documentada na literatura científica.

- O volume, mais expressivo dos três, refere ao tamanho dos dados. A enorme quantidade de dados e bases distintas que estamos lidando atualmente exigem que os cientistas repensem os paradigmas

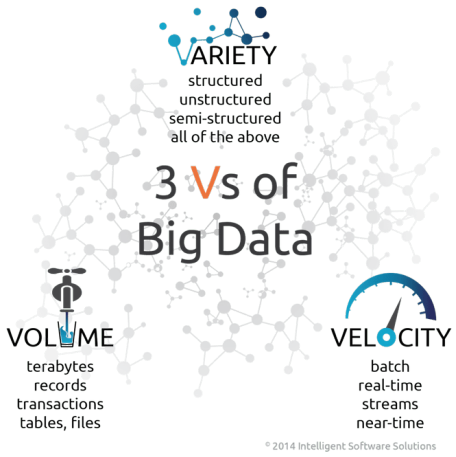



Figura 1. Os 3V's do Big Data

correntes de armazenamento e processamento, a fim de desenvolver as ferramentas necessárias para analisar os dados de forma adequada e eficiente.

- A velocidade aborda o tempo que os dados podem ser recebidos e analisados. Essa dimensão do problema é particularmente notável em tarefas como processamento de *streams*, que analisam os dados em tempo real, à medida em que são gerados.
- Variedade refere-se à questão de formatos de dados diferentes e incompatíveis. Os dados podem vir de várias fontes e assumir muitas formas diferentes, e apenas prepará-los para análise leva uma quantidade significativa de tempo e esforço.

Atualmente, o problema das grandes coleções de dados é, muitas vezes, resolvido através de sistemas de armazenamento distribuídos, que são projetados para controlar cuidadosamente o acesso e prover um sistema de gerenciamento que faz com que o sistema seja tolerante a falhas. Uma solução para o problema dos grandes objetos de dados na aprendizagem de máquinas é através da paralelização de algoritmos. Isso geralmente é realizado de duas maneiras [3]: paralelismo de dados, no qual os dados são divididos em partes gerenciáveis em uma só máquina e cada subconjunto é computado simultaneamente, ou o paralelismo de tarefas, no qual o algoritmo é dividido em etapas que podem ser realizadas simultaneamente.

À medida que os dados crescem e se tornam mais amplamente disponíveis, é cada vez mais comum a necessidade de tratá-los e obter, a partir deles, informações relevantes a respeito do assunto tratado. O aumento na facilidade ao acesso a um poder de computação cada vez maior (seja através de máquinas com alto desempenho, seja através de serviços em nuvem), aliado ao crescimento vertiginoso da disponibilidade dos dados está abrindo muitas novas oportunidades para pesquisas em aprendizado de máquina. Muitas dessas novas direções utilizam fluxos de trabalho cada vez mais complexos, que exigem sistemas construídos usando uma combinação de ferramentas e técnicas de última geração.

Os métodos tradicionais de aprendizado de máquina automático realizam o desenvolvimento de modelos off-line ou batch em um conjunto de dados para treinamento antes de considerar novos dados para teste. Essa abordagem é útil quando há uma massa de dados históricos disponíveis para alimentar uma modelo de aprendizado de máquina. Em algumas situações, porém, não existem dados históricos disponíveis, sendo necessário utilizar um modelo online para executar a tarefa desejada. Segundo a abordagem online, os modelos de ML são construídos e avaliados à medida que os dados entram no sistema, exigindo algoritmos eficientes que possam processar os dados na velocidade em que eles estão chegando. Os algoritmos de aprendizado de máquina são bastante variados, e cada um é indicado para uma tarefa específica. Abaixo estão algumas famílias de algoritmos e uma breve introdução a cada grupo, a figura 2 [4] pode ajudar a elucidar:

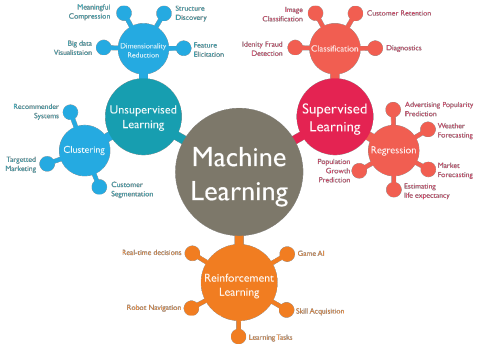


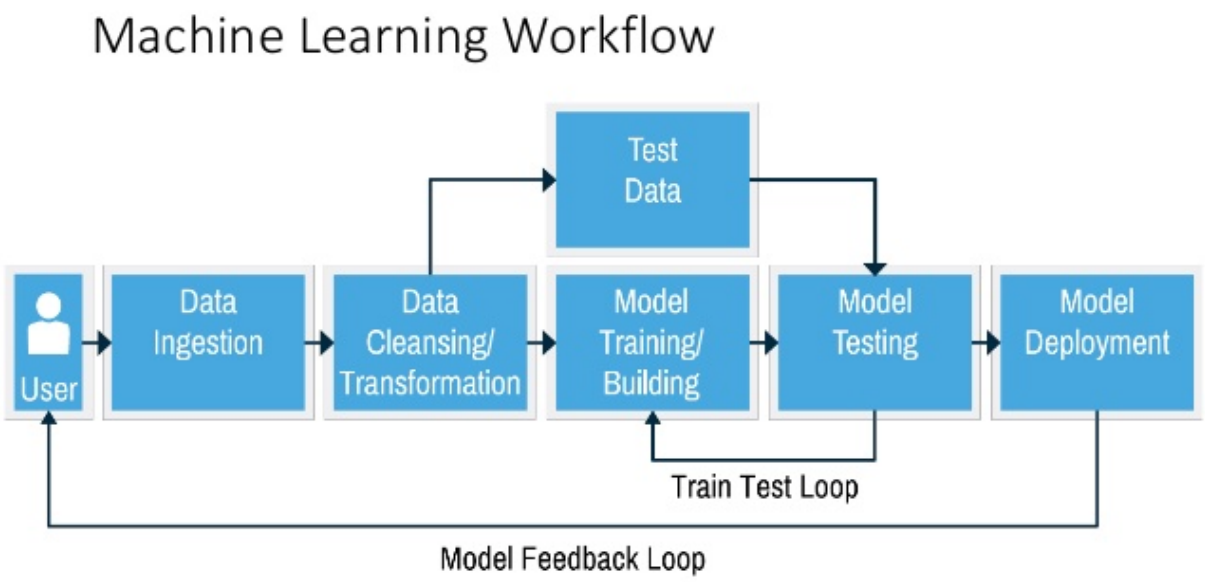

Figura 2. Categorias do Machine Learning

- Classificação: O processo de classificação é uma tarefa de aprendizagem supervisionada que discrimina um conjunto de objetos ou dados por rótulos de classe pré-determinados. Alguns exemplos básicos são rotulagem de e-mail como spam e análise do sentimento dos dados de texto [5]. Algoritmos usados para classificação são Regressão Logística, Árvores de Decisão, Máquinas de Vetor de Suporte e descriminação de k-vizinhos mais próximos.
- Clustering: Clustering é uma tarefa de aprendizagem sem supervisão que agrupa elementos em uma população ao examinar suas diversas características. Os algoritmos comuns de agrupamento são clustering de k-Means e clusterização de expectativa-maximização (EM).
- Regressão: A regressão também é uma tarefa de aprendizagem supervisionada que adapta um modelo a um grupo de dados, realizando previsões numéricas acerca de informações em comum [6]. Os algoritmos tradicionais para regressão incluem regressão linear e modelos de árvores.
- Recomendação: os sistemas de recomendação preveem relações significativas entre entidades utilizando interações e comportamentos de usuários e

com isso procuram itens que se assemelhem às preferências dos mesmos. Recomendar livros, músicas, filmes e outros produtos aos usuários com base em seus relacionamentos com outros usuários é um excelente exemplo desses sistemas. Os métodos de filtragem colaborativa (CF) são utilizados para criar sistemas recomendadores [7] e evoluíram muito, principalmente em sites de comércio eletrônico, como a Amazon, onde esse tipo de serviço é essencial [8].

- Deep Learning: utiliza Redes Neurais que simulam o cérebro humano para modelar relações matemáticas entre os dados. O Deep Learning é um campo promissor de aprendizagem de máquinas e inteligência artificial [9], e tem sido usado para detectar objetos em vídeos e vencer jogadores humanos em videogames [10]. Embora o Deep Learning seja usado para tarefas como classificação e regressão, consideramos ela uma categoria separada das outras para ilustrar a cobertura pelas várias ferramentas e aplicabilidade em diversos cenários.
- Regras de Associação: Os modelos dessa categoria, como Frequent Pattern Growth (FP-Growth), realizam a detecção de relações entre objetos em um conjunto de dados. Essas regras são usadas para destacar características que não estão preenchidas, não existam ou selecionar conjuntos de itens que estão relacionados entre si.
- Redução de Dimensionalidade: Algoritmos dessa natureza reduzem o tamanho dos dados combinando, transformando e removendo recursos ou características que são apontados como pouco relevantes em uma modelagem descritiva / objetiva dos dados. Algoritmos notáveis para redução de dimensionalidade são a Análise de Componentes Principais (PCA) e Decomposição em Valores Singulares (SVD).

[https://medium.com/@agarwalvibhor84/getting-started-with-machine-learning-using-sklearn-python-7d165618eddf]




Esses algoritmos de aprendizado de máquina são frequentemente utilizados para extrair informação útil de grandes conjuntos de dados. Essa tarefa, conhecida como Mineração de Dados, porém, engloba um contexto bem mais amplo e que merece um detalhamento maior, que é fornecido a seguir.

## 3. Mineração de Dados

A definição mais aceita do conceito de Mineração de Dados, ou *Data Mining* é a descoberta de modelos para um determinado conjunto de dados. O conceito de "modelo", porém, é bastante amplo e, estreitar a definição desse termo é um passo absolutamente necessário para que se possa obter uma ideia significativa a respeito do real significado desse conceito [11]. Considerando este contexto, podemos fazer uma distinção do termo "modelo" em 3 categorias:

- Modelagem Estatística: Os profissionais da área da Estatística foram os primeiros a utilizar o termo "Data Mining" (Mineração de Dados), que à época tinha uma conotação bastante negativa, referindo-se a tentativas de extrair informações de dados que não davam suporte ao que pretendia-se concluir. Nos dias de hoje, a expressão Data Mining tem um significado

mais positivo e, para os estatísticos, significa construir um modelo representativo da distribuição de onde foram retirados os dados. Um exemplo simples para se entender esta modelagem é pensar em um dataset composto apenas por uma coleção de números. Neste caso, de acordo com a abordagem estatística, a mineração de dados consistiria em determinar o tipo de distribuição que deu origem aos dados e descrevê-la. Em uma distribuição Gaussiana, por exemplo, o modelo consistiria da média e do desvio padrão, já que essas informações são suficientes para descrever como os dados se comportam.

- Modelagem baseada em aprendizado de máquina: Alguns estudiosos da área consideram os termos "mineração de dados" e "aprendizado de máquina" como sendo sinônimos e é perfeitamente possível compreender os motivos para tal, uma vez que ambos os termos estão fortemente relacionados. A abordagem baseada em aprendizado de máquina consiste em utilizar um subconjunto dos dados para treinar determinados algoritmos e utilizar os produtos destes treinamentos para prever o comportamento de registros de dados ainda não vistos. Em alguns cenários, utilizar os dados dessa maneira faz sentido. Normalmente, aprendizado de máquina é uma boa abordagem quando não se sabe exatamente o que procurar nos dados. Um bom exemplo é o caso de recomendação de filmes, por exemplo, já que os motivos que fazem com que uma determinada pessoa goste ou não de um filme não são muito claros de um ponto de vista objetivo. Em um caso como esse, a análise de uma amostra de avaliações de filmes previamente assistidos se mostra bastante eficiente na tarefa de prever as avaliações de filmes ainda não assistidos. Por outro lado, existem casos em que esta abordagem não se mostra eficiente. Tipicamente, são casos em que se pode descrever os objetivos da mineração de maneira mais direta. Um caso exemplo dessa situação foi a tentativa da empresa WhizBang! Labs de utilizar aprendizado de máquina para tentar identificar páginas na internet que continham currículos de pessoas. O desempenho dos algoritmos não se mostrou superior ao de outros, implementados diretamente, que utilizavam palavras óbvias (cuja presença em currículos era esperada) para fazer essa identificação.
- Modelagem baseada em abordagens computacionais: Neste caso, o modelo dos dados é simplesmente a resposta para uma pergunta complexa a respeito do dataset. A maioria das abordagens computacionais podem ser divididas em duas categorias:

    1) Resumir os dados sucintamente: Um dos exemplos dessa proposta é o algoritmo de PageRank, que resume a estrutura da web em um conjunto de números, cada um associado a uma página. Este número, conhecido como o PageRank da página indica a probabilidade de um random walker caminhando no grafo da web esteja naquela página em qualquer instante de tempo. A propriedade interessante desse sistema de ranqueamento é que o número associado a cada página reflete muito bem a "importância" daquela página, isto é, o quanto os usuários gostariam que aquela página aparecesse em suas buscas.
    2) Extrair características: Um modelo baseado em características aborda os dados através da análise dos registros mais extremos do conjunto. Alguns dos mais importantes tipos extração de características comumente utilizados em dados massivos são:

        a) Itens frequentes: Esse modelo é aplicado quando os dados consistem de pequenos conjuntos de itens de uma determinada coleção. O objetivo é encontrar itens que frequentemente aparecem juntos nesses conjuntos. Aplicando um modelo como esse em um conjunto de dados de cestas de supermercados, por exemplo, é provável que encontremos hambúrgueres e ketchup como pares frequentes, uma vez que são alimentos frequentemente consumidos (e comprados) juntos.
        b) Itens similares: Esse modelo é aplicado quando os dados são uma coleção de conjuntos e o objetivo é encontrar pares de conjuntos que tem uma grande parcela de elementos em comum. Um exemplo de aplicação de um algoritmo como esse é o processo de collaborative filtering, muito utilizado em sites de comércio eletrônico, que consiste em detectar clientes que têm interesse em muitos itens em comum e recomendar a cada um deles os itens que não são comuns. A premissa é que se ambos os clientes já possuem interesse em uma grande quantidade de produtos, é provável que eles venham a se interessar por produtos que chamaram a atenção um do outro.

Dito todas essas categorias vemos a relação que área podem cooperar para aperfeiçoamentos das técnicas utilizadas e cenários aplicáveis e com isso algumas ferramentas que iremos abordar na próxima seção são eles: TensorFlow do Google, MLIB, Mahout e Azure Machine Learning da Microsoft.

## 4. Ferramentas e Tecnologias

### 4.1. TensorFlow

TensorFlow™ [1] é uma biblioteca de software de código aberto para computação numérica usando grafos de fluxo de dados. Os nós do grafo representam operações matemáticas, enquanto suas arestas representam fluxo de dados multidimensionais (tensors). Por possuir uma arquitetura flexível, permite distribuir computação para uma ou mais CPUs ou GPUs em uma estação de trabalho, servidor ou dispositivo móvel com uma única API. O TensorFlow foi originalmente desenvolvido por pesquisadores e engenheiros que trabalham no Google Brain Team dentro da organização de pesquisa de Machine Intelligence do Google para fins de pesquisa em aprendizado de máquinas e redes neurais profundas. Deve-se ressaltar, no entanto, que o sistema é flexível o suficiente para ser aplicável também em uma variedade de outros domínios. [12]

Atualmente na versão 1.4, o TensorFlow é a biblioteca mais utilizada no mercado para desenvolvimento de aplicações que utilizam aprendizado de máquinas distribuído. A plataforma foi criada para ser a sucessora do DistBelief, que é um sistema para o treinamento de redes neurais que o Google vem utilizando desde 2011. O TensorFlow runtime é uma biblioteca de várias plataformas. A figura 3 [12] ilustra sua arquitetura: uma API escrita em C separa o código do nível do usuário em diferentes camadas, o tempo de execução do núcleo de processamento. A biblioteca central TensorFlow é implementada em C ++, priorizando portabilidade e desempenho. A ferramenta pode ser executada com performance satisfatória em várias plataformas de hardware e software, como Linux, OS X, Windows, Android e iOS; rodando tanto sobre arquitetura x86, quanto sobre outras soluções baseadas em ARM. Também é possível utilizar arquiteturas de GPU projetadas pela NVIDIA, que realiza esforços no sentido de oferecer processadores massivamente paralelos para aprendizado de máquina. A implementação é de código aberto e aceita várias contribuições externas que permitem que o TensorFlow seja executado em outras arquiteturas.

A API do TensorFlow é dividida em camadas que desempenham papéis únicos para aprendizagem de máquina e integração de fluxos de execução paralelos. O fluxo de dados simplifica a execução distribuída, uma vez que torna a comunicação entre subcomputações mais explícita. Ele permite que o mesmo programa TensorFlow seja implantado em um cluster de GPUs para treinamento, um conjunto de TPUs(Tensor Processing Unit) para processamento e um celular para gerenciamento móvel, por exemplo. Cada operação é computada em um dispositivo determinado, cada CPU ou GPU executa uma tarefa específica. Um dispositivo é responsável por executar um kernel para cada operação atribuída a ele e a biblioteca permite que vários kernels sejam registrados para uma única operação, fornecendo

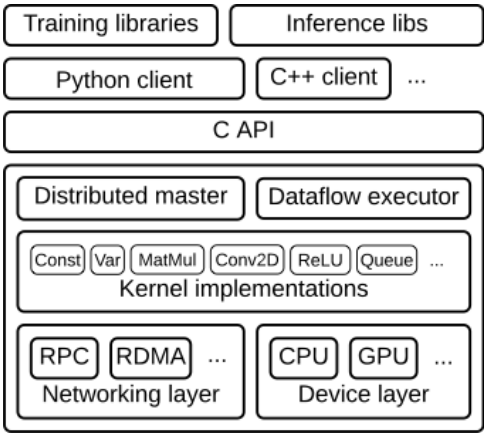



Figura 3. Arquitetura TensorFlow

implementações especializadas para um determinado dispositivo ou tipo de dados. A Figura 4 [12] demonstra as interconexões das diversas camadas da API. Algumas das aplicações modeladas segundo a proposta do TensorFlow dentro do próprio ambiente do Google são a categorização de imagens,e a recomendação de palavras a serem inseridas a medida que um texto é escrito [12].

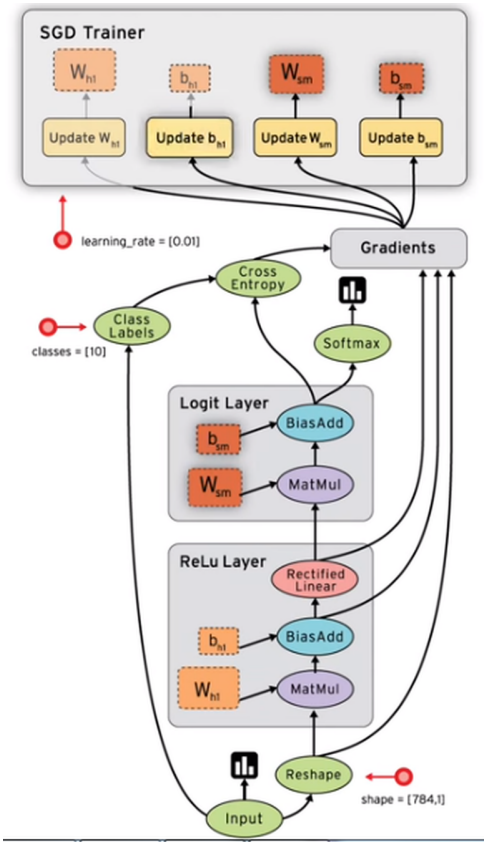



Figura 4. camadas TensorFlow

1. https://www.tensorflow.org/

### 4.2. SciKit-learn

O Scikit-learn aproveita o rico ambiente do Python para fornecer aos desenvolvedores de aplicações, implementações de muitos algoritmos de aprendizado de máquinas bem conhecidos, mantendo uma interface fácil de usar e muito bem integrada com o restante da linguagem. A biblioteca almeja responder à crescente demanda de análise de dados estatísticos por não especialistas em software e indústrias da web, bem como em áreas fora da informática, como biologia ou física. A biblioteca foi projetada para se conectar ao conjunto de pacotes numéricos e científicos centrados nas bibliotecas NumPy e SciPy. NumPy [13] acrescenta ao Python um tipo de dados numéricos contíguos e primitivas de computação de matriz eficientes, enquanto SciPy estende ainda mais com operações numéricas comuns, seja implementando estas em Python/NumPy ou envolvendo implementações C/C++/Fortran existentes. Com base nesse *stack*, foram criadas uma série de bibliotecas denominadas scikits, para complementar o SciPy com kits de ferramentas específicos de diversos domínios científicos. Atualmente, os dois mais populares e completos são, de longe, scikit-learn e o scikit-image.

Todos os objetos dentro do scikit-learn compartilham uma API comum, uniforme e composta de três interfaces complementares: uma interface estimadora para construção de modelos e montagem, uma interface para fazer previsões e uma interface para transformar e converter dados. A interface do estimador está no centro da biblioteca. Ela define mecanismos de instanciação de objetos e expõe um método adequado para que a máquina aprenda um modelo a partir de dados de treinamento. Todos os algoritmos de aprendizagem supervisionados e não supervisionados são oferecidos como objetos que implementam essa interface. Alguns exemplos de tarefas de que são fornecidos são seleção e extração de recursos, e redução de dimensionalidade.

O SciKit-learn, codifica os dados de maneira mais próxima possível à representação de uma matriz. Os conjuntos de dados são codificados como matrizes multidimensionais NumPy para dados densos e como matrizes SciPy para dados esparsos. Embora tais codificações possam parecer representações de dados pouco sofisticadas quando comparadas a construções orientadas a objetos, como as usadas pelo Weka, elas trazem a vantagem de permitirem ao desenvolvedor contar com operações vetorizadas NumPy e SciPy eficientes, enquanto mantém o código curto e legível. Do ponto de vista prático, esses formatos também fornecem uma coleção de ferramentas de carregamento e conversão de dados que os tornam muito fáceis de usar em muitos contextos. Além disso, para tarefas em que as entradas são arquivos de texto ou objetos semi-estruturados, são fornecidos objetos vetorizados que convertem eficientemente esses dados nos formatos esperados pelas bibliotecas subjacentes [14].

A Figura 5 [15], demonstra como o SciKit learn direciona a modelagem de um problema de aprendizagem de máquina dentre as soluções implementadas pela plataforma e como essas soluções se encaixam dentre as categorias de algoritmos apresentadas na Seção 2. É interessante notar que a biblioteca vem sendo utilizada em diversas áreas e setores da indústria, e, uma aplicação que merece destaque é a classificação de neuroimagens para geração de determinados diagnósticos médicos de forma automática [16].

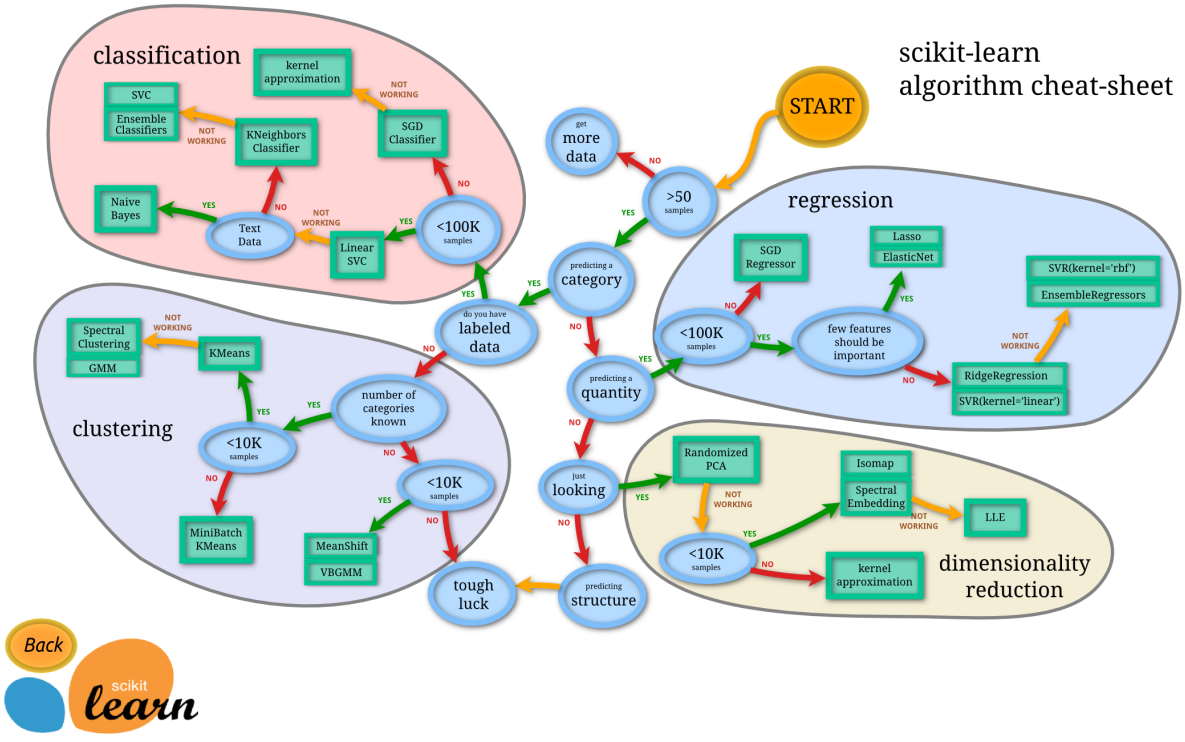



Figura 5

### 4.3. Mahout, MLlib e SAMOA

Essas três bibliotecas são amplamente utilizadas para a criação de sistemas de aprendizado de máquina em larga escala, muitas vezes fazendo uso de motores de processamento distribuído. Tais soluções são frequentemente comparadas e avaliadas em *surveys* [17] [18], e algumas de suas principais características estão compiladas a seguir:

MLlib[2] é uma biblioteca de aprendizado de máquina que funciona através da plataforma Spark, um mecanismo de processamento distribuído em memória de dados que rapidamente ganhou popularidade e vem sendo adotado para tarefas que fazem uso de dados massivos. O Spark suporta o processamento de batch e stream, e o MLlib pode ser usado para aprender com dados usando ambos os paradigmas.

O Mahout[3] é a ferramenta mais antiga para a aprendizagem de máquinas distribuídas. Inicialmente, a biblioteca foi construída com base em Hadoop e MapReduce como uma estrutura de aprendizado de máquina em batch e foi bastante utilizada tanto na academia quanto no mundo corporativo, aproveitando-se da popularidade do Hadoop.

Já o SAMOA[4] (Scalable Advanced Massive Online Analysis) foi desenvolvido pelo Yahoo! Labs especificamente para aprendizagem de máquinas a partir de streams de dados. O projeto esteve em incubação do projeto Apache desde o final de 2014. A plataforma fornece construção e avaliação on-line em tempo real de modelos de aprendizagem de máquinas a partir de streams de dados de diversas fontes. O projeto foi concebido com o objetivo de suprir a falta da capacidade de lidar com streams do Mahout e possui, portanto, semelhanças consideráveis com o sistema que o originou.

## 5. Aplicações

Machine Learning é uma *buzzword* no mundo da tecnologia no momento. Isso ocorre pelo fato de que o desenvolvimento deste campo de pesquisa representa um grande passo adiante no modo como os computadores podem processar e analisar dados de maneira mais autônoma, dependendo cada vez menos de interferência humana. Como foi visto nas seções anteriores, existem diversos modelos e categorias de algoritmos de aprendizado de máquina, e seu uso vem resultando no desenvolvimento de uma variedade de aplicações práticas que têm impacto real na vida das pessoas. Alguns exemplos de aplicações desta tecnologia são:

1) Segurança de dados: No ano de 2014, a Kaspersky Labs afirmou ter detectado 325 mil novos arquivos de malware todos os dias. É importante notar, porém, que de acordo com a empresa de inteligência computacional "Deep Instinct", cada novo malware tende a ter quase o mesmo código que as versões anteriores - apenas entre 2 e 10% dos arquivos mudam de iteração para iteração [5]. Seu modelo de aprendizagem não tem nenhum problema com as variações de 2-10% e pode prever quais arquivos são malwares com grande precisão. Em outras situações, os algoritmos de aprendizagem de máquinas podem procurar padrões em como os dados na nuvem são acessados e relatam anomalias que poderiam prever brechas de segurança. [19]
2) Segurança pessoal: A aprendizagem de máquinas está provando que pode ser uma ótima ferramenta para ajudar a eliminar falsos alarmes e a detectar ameaças que os humanos podem deixar passar em testes de segurança em aeroportos, estádios, concertos e outros locais com grandes concentrações de pessoas. Isso pode acelerar o processo de checagem significativamente e garantir eventos mais seguros. [20]
3) Mercado financeiro: A capacidade de prever o comportamento do mercado financeiro com certa precisão é uma fator fundamental para o sucesso de investidores, acionistas e empresários. Muitas empresas comerciais de prestígio utilizam sistemas proprietários baseados em aprendizagem de máquina para prever as tendências do mercado e assim, obter vantagens competitivas ao executar negócios de grandes proporções. [21] [22]
4) Cuidados com a saúde: Os algoritmos de aprendizagem de máquina podem processar mais informações e detectar mais padrões do que os seus homólogos humanos. Um estudo utilizou o diagnóstico assistido por computador (CAD) para analisar exames de mamografia de mulheres que mais tarde, viriam a desenvolver câncer de mama. Os resultados obtidos pelo computador foram capazes de detectar 52% dos casos até um ano antes das mulheres serem oficialmente diagnosticadas [23]. Um outro exemplo de aplicação de algoritmos de aprendizado de máquina na área da saúde é o caso da empresa Medecision [6], que desenvolveu uma ferramenta capaz de identificar oito variáveis para prever hospitalizações evitáveis em pacientes com diabetes.
5) Personalização de Marketing: Em muitas lojas virtuais, algoritmos de aprendizado de máquina são utilizados para recomendar produtos aos clientes baseados no comportamento dos mesmos em momentos anteriores. Normalmente, produtos parecidos com ou relacionados a outros previamente adquiridos são recomendados. Baseadas em comportamentos prévios dos usuários, as empresas podem personalizar e-mails contendo propaganda, oferecer promoções exclusivas, manipular quais produtos

2. https://spark.apache.org/mllib/
3. http://mahout.apache.org/
4. https://samoa.incubator.apache.org/
5. https://www.blackhat.com/docs/us-16/materials/us-16-Nipravsky-Certificate-Bypass-Hiding-And-Executing-Malware-From-A-Digitally-Signed-Executable-wp.pdf
6. https://www.medecision.com/

são exibidos primeiro no site, etc. Essa questão levanta uma série de outros pontos que merecem discussão, e por isso, a seção 5.1 traz uma abordagem mais detalhada a respeito de temas correlatos.

6) Detecção de fraude: A aprendizagem de máquinas está melhorando a cada dia a detecção de possíveis casos de fraude em vários campos diferentes. O PayPal, por exemplo, vem usando *machine learning* para combater o desvio de capitais. A empresa possui ferramentas que comparam milhões de transações e podem distinguir precisamente entre transações legítimas e fraudulentas entre compradores e vendedores. [24]
7) Recomendação de Conteúdo: Serviços como Spotify e Netflix possuem sistemas complexos de recomendação de conteúdo e grande parte de seu sucesso pode ser atribuído à precisão alcançada por esses algoritmos na tarefa de decidir quais itens da biblioteca tem mais chances de agradar os usuários. Os algoritmos inteligentes de aprendizagem de máquinas analisam a atividade de um determinado cliente e as comparam com as de milhões de outros usuários para determinar o que aquele indivíduo pode estar interessado. [25]
8) Pesquisas Online: Empresas como o Google e seus concorrentes estão melhorando constantemente o que o mecanismo de busca compreende. Toda vez que um usuário executa uma pesquisa no Google, por exemplo, o programa observa como o indivíduo responde aos resultados retornando e utiliza essas reações para melhorar os resultados de buscas futuras. [21]
9) Processamento da linguagem natural (PNL): O PNL está sendo usado em todo tipo de aplicações, em uma variada gama de disciplinas. Os algoritmos de aprendizado de máquina para lidar com linguagem natural são comumente utilizados em aplicações que automatizam processos de atendimento ao cliente, podendo inclusive ajudar a direcioná-los mais rapidamente para a informação de que precisam. Outras aplicações são a detecção de cláusulas obscuras em contratos, o que pode ajudar advogados a selecionar quais dados são cabíveis de serem utilizados em um processo. [26]
10) Smart Cars: A IBM recentemente fez uma pesquisa com os principais produtores de automóveis e concluiu que 74% deles esperavam que veríamos carros inteligentes nas estradas até 2025[7]. Um carro inteligente não só seria capaz de se conectar à Internet, mas também aprenderia informações sobre seu proprietário e seu ambiente. Através desses dados, o carro seria capaz de ajustar as configurações internas - temperatura, áudio, posição do assento, etc. - automaticamente, gerar relatórios, dirigir sozinho e oferecer dados em tempo real sobre o tráfego e as condições da estrada. [27]

7. http://www-935.ibm.com/services/multimedia/GBE03640USEN.pdf

## 5.1. Anúncios Online

Uma das grandes novidades trazidas com a chegada do século XXI e da era de uma sociedade massivamente conectada foi o advento de uma infinidade de aplicações web gratuitas para o usuário e que os oferecem uma grande variedade de comodidades. Esse fenômeno se deve ao fato de que as empresas e organizações responsáveis pelo desenvolvimento e manutenção desses serviços tornaram-se capazes de sustentar seus negócios através da exibição eficiente de anúncios aos seus usuários. O mercado de publicidade online logo se tornou um negócio bilionário, dando origem a grandes companhias de tecnologia, como o Google, o Yahoo e o eBay. O sucesso dessas grandes organizações se deve basicamente ao seu potencial de conseguir entregar soluções publicitárias aos anunciantes de uma forma muito mais eficiente do que os veículos midiáticos tradicionais.

Um anúncio de um carro, por exemplo, publicado em um jornal que aborda temas diversificados, como a Folha de São Paulo, tem uma efetividade bem limitada. Apesar do número alto de pessoas que verão a propaganda, é bastante provável que muitas delas não tenham interesse em comprar um carro no momento, não tenham habilitação para dirigir, tenham acabado de comprar um outro veículo ou tenham algum outro motivo para ignorar o anúncio. Mesmo assim, o custo da impressão do anúncio naquele exemplar do jornal comprado por estes leitores teve que ser pago pelo anunciante, o que é uma maneira ineficiente de investir recursos. O mesmo anúncio publicado em um grande portal da Internet sofre do mesmo problema, com a diferença de que o custo de publicação do anúncio na página online é consideravelmente menor do que no caso do jornal. A alternativa da mídia tradicional para tratar esse problema foi a criação de revistas e jornais que tratam de temas específicos. Um anúncio de um taco de golfe em uma revista especializada nesse esporte tem muito mais chances de não ser ignorado por um leitor do que o mesmo anúncio publicado em uma revista mais generalista. O mesmo fenômeno acontece na Internet, e as chances de um anúncio receber um clique são muito maiores se o mesmo for publicado em um site especializado em assuntos relacionados ao produto. O ambiente online possui, porém, uma vantagem considerável com relação às mídias físicas: é possível usar informações a respeito do usuário para decidir, em tempo real, qual anúncio exibir a ele, independentemente da página que está sendo acessada. Se o fato de que Lucy gosta de golfe é conhecido, então faz sentido mostrá-la anúncios relacionados a golfe independentemente da página que ela está acessando. Se é possível conhecer quais páginas Lucy acessa, quanto tempo ela passa em cada uma, seu histórico de mensagens de e-mail, quais buscas ela faz em sites como o Google e quais grupos ela participa no Facebook, por exemplo, então é possível utilizar técnicas e ferramentas de mineração de dados para descobrir, indiretamente, quais são seus gostos pessoais e, com isso, exibir anúncios mais relevantes a ela. Não é difícil perceber, contudo, que práticas como essas levantam inúmeras questões de privacidade que, na prática, não têm soluções que satisfaçam todas as preocupações inerentes ao tema. Por

um lado, as pessoas querem utilizar serviços web gratuitos e de alta qualidade, mas por outro, a simples existência desses serviços depende da sua capacidade de fornecer a anunciantes soluções publicitárias muito mais eficazes do que a mídia tradicional. Essa capacidade, contudo, depende do conhecimento que a empresa tem sobre seu público e, portanto, a privacidade nem sempre pode ser preservada. Estas questões levantam também uma série de grandes preocupações que são de extrema relevância neste meio. Existe o risco de que os gostos e preferências dos usuários deixem de ser usados apenas para fins publicitários e passem a ser ferramentas nas mãos de pessoas mal-intencionadas. Governos autoritários podem ser capazes de identificar, com extrema facilidade, inimigos políticos, scammers podem ser capazes de obter informações detalhadas a respeito de suas vítimas, empresas podem ser capazes de investigar a vida de seus funcionários e tirar proveito dessas informações, etc. Acredita-se, portanto, que pesquisas ainda devem ser feitas na área no sentido de evitar que a mineração de dados venha agravar ainda mais os problemas já inerentes a uma sociedade massivamente conectada.

## 6. Conclusão

Em vista do tema apresentado e das discussões propostas, fica fácil observar a grande importância e o crescimento de três grandes áreas da ciência da computação: *Big Data, Machine Learning* e mineração e dados. São áreas extremamente relacionadas e a exploração de suas sobreposições tem rendido frutos práticos e teóricos não apenas à computação como ciência, mas também a toda a Tecnologia da Informação como ferramenta de transformação social. O objetivo deste trabalho era fazer uma breve abordagem, realizando uma seleção de artigos, livros e referências na internet que ajudem a elucidar esse amplo cenário que emerge, cada vez mais rápido, no futuro da tecnologia e da sociedade como um todo. Acredita-se que os objetivos tenham sido cumpridos e que o trabalho concluído seja uma colaboração válida no sentido de introduzir o tema a um leitor leigo e provocar discussões entre leitores mais avançados. É evidente que mais contribuições devem ser feitas na área, especialmente no que tange aos desafios tecnológicos e sociais que despontam como barreiras dentro deste contexto. É importante ressaltar que as tecnologias de aprendizado de máquina, *big data* e mineração de dados estão transformando o modo como o homem lida com a computação e tem portanto, enorme potencial para transformar o modo como a sociedade funciona. É nossa responsabilidade como cientistas, portanto, comprometermo-nos com o desenvolvimento de técnicas e ferramentas que venham trazer impacto positivo na vida das pessoas, contribuindo assim para a construção de um futuro que valha a pena.

## Referências


[1] D. Laney, "3d data management: Controlling data volume, velocity and variety," *META Group Research Note*, vol. 6, p. 70, 2001.

[2] M. Z. A. Taie. (2017) os 3vs do bg data. [Online]. Available: http://blog.agroknow.com/?p=3667

[3] R. Bekkerman, M. Bilenko, and J. Langford, *Scaling up machine learning: Parallel and distributed approaches*. Cambridge University Press, 2011.

[4] V. Agarwal. (2017) As categorias de machine learning. [Online]. Available: https://medium.com/@agarwalvibhor84/getting-started-with-machine-learning-using-sklearn-python-7d165618eddf

[5] T. M. Khoshgoftaar and N. Seliya, "Fault prediction modeling for software quality estimation: Comparing commonly used techniques," *Empirical Software Engineering*, vol. 8, no. 3, pp. 255–283, 2003.

[6] J. Van Hulse and T. M. Khoshgoftaar, "Class noise detection using frequent itemsets," *Intelligent Data Analysis*, vol. 10, no. 6, pp. 487–507, 2006.

[7] X. Su and T. M. Khoshgoftaar, "A survey of collaborative filtering techniques," *Adv. in Artif. Intell.*, vol. 2009, pp. 4:2–4:2, Jan. 2009. [Online]. Available: http://dx.doi.org/10.1155/2009/421425

[8] B. Smith and G. Linden, "Two decades of recommender systems at amazon.com," *IEEE Internet Computing*, vol. 21, no. 3, pp. 12–18, May 2017.

[9] M. M. Najafabadi, F. Villanustre, T. M. Khoshgoftaar, N. Seliya, R. Wald, and E. Muharemagic, "Deep learning applications and challenges in big data analytics," *Journal of Big Data*, vol. 2, no. 1, p. 1, 2015.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. A. Riedmiller, "Playing atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013. [Online]. Available: http://arxiv.org/abs/1312.5602

[11] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of massive datasets*. Cambridge university press, 2014.

[12] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. J. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Józefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. G. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. A. Tucker, V. Vanhoucke, V. Vasudevan, F. B. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *CoRR*, vol. abs/1603.04467, 2016. [Online]. Available: http://arxiv.org/abs/1603.04467

[13] S. v. d. Walt, S. C. Colbert, and G. Varoquaux, "The numpy array: a structure for efficient numerical computation," *Computing in Science & Engineering*, vol. 13, no. 2, pp. 22–30, 2011.

[14] L. Buitinck, G. Louppe, M. Blondel, F. Pedregosa, A. Mueller, O. Grisel, V. Niculae, P. Prettenhofer, A. Gramfort, J. Grobler, R. Layton, J. VanderPlas, A. Joly, B. Holt, and G. Varoquaux, "API design for machine learning software: experiences from the scikit-learn project," *CoRR*, vol. abs/1309.0238, 2013. [Online]. Available: http://arxiv.org/abs/1309.0238

[15] X.Ren. (2017) Python data engineering cheat sheet. [Online]. Available: http://xren615.github.io/post/cheatSheet/

[16] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux, "Machine learning for neuroimaging with scikit-learn," *Frontiers in neuroinformatics*, vol. 8, 2014.

[17] S. Landset, T. M. Khoshgoftaar, A. N. Richter, and T. Hasanin, "A survey of open source tools for machine learning with big data in the hadoop ecosystem," *Journal of Big Data*, vol. 2, no. 1, p. 24, 2015.

[18] A. N. Richter, T. M. Khoshgoftaar, S. Landset, and T. Hasanin, "A multi-dimensional comparison of toolkits for machine learning with big data," in *Information Reuse and Integration (IRI), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–8.

[19] D. Zhu, H. Jin, Y. Yang, D. Wu, and W. Chen, "Deepflow: Deep learning-based malware detection by mining android application for abnormal usage of sensitive data," in *Computers and Communications (ISCC), 2017 IEEE Symposium on*. IEEE, 2017, pp. 438–443.

[20] N. Zhang and J. Zhu, "A study of x-ray machine image local semantic features extraction model based on bag-of-words for airport security." *BioTechnology: An Indian Journal*, vol. 10, no. 24, 2014.

[21] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[22] J. Patel, S. Shah, P. Thakkar, and K. Kotecha, "Predicting stock and stock price index movement using trend deterministic data preparation and machine learning techniques," *Expert Systems with Applications*, vol. 42, no. 1, pp. 259–268, 2015.

[23] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Computational and structural biotechnology journal*, vol. 13, pp. 8–17, 2015.

[24] J. A. Gómez, J. Arévalo, R. Paredes, and J. Nin, "End-to-end neural network architecture for fraud scoring in card payments," *Pattern Recognition Letters*, 2017.

[25] C. A. Gomez-Uribe and N. Hunt, "The netflix recommender system: Algorithms, business value, and innovation," *ACM Transactions on Management Information Systems (TMIS)*, vol. 6, no. 4, p. 13, 2016.

[26] H. Surden, "Machine learning and law," 2014.

[27] M. Whaiduzzaman, M. Sookhak, A. Gani, and R. Buyya, "A survey on vehicular cloud computing," *Journal of Network and Computer Applications*, vol. 40, pp. 325–344, 2014.