

MACHINE LEARNING NO MUNDO REAL

ESTUDOS DE CASO, TÉCNICAS E RISCOS

Agosto 2019



InfoQ

ARTIGO [pág 6](#)

Mais bytes
no seu
bolso

ARTIGO [pág 13](#)

De volta para o Futuro:
desmistificando o viés
cognitivo

ARTIGO [pág 24](#)

Analisando e prevenindo o
preconceito inconsciente
em Machine Learning

NESTA EDIÇÃO

6

Mais bytes no seu bolso

A Lovethesales.com teve que classificar dados de um milhão de produtos de mais de 700 fontes diferentes em um vasto domínio. Eles decidiram criar uma hierarquia de classificadores utilizando machine learning, especificamente Support Vector Machines. Eles aprenderam que otimizando a maneira em que svms foram conectadas levando a várias melhorias no reuso de dados de treinamento rotulados.

13

De volta para o Futuro: Desmistificando o Viés Cognitivo

A IA nas empresas tem nuances mais predominantes nos dados de entrada quando comparado com a IA aplicada a um consumidor ou na academia. O calcanhar de Aquiles neste domínio é o viés cognitivo. Em termos leigos, é como o Marty McFly (De volta para o Futuro) viajando para o futuro, colocando as mãos no almanaque esportivo e usando-o para apostar nos jogos do presente. Mayukh Bhaowal, do Salesforce Einstein, explica como neutralizar este efeito.

18

Entendendo o comportamento de sistemas e softwares com Machine Learning e dados de séries temporais

No QCon.ai 2018, David Andrzejewski apresentou "Entendendo o Comportamento Sistêmico de Softwares com Machine Learning e dados de séries temporais". David é gerente de engenharia na Sumo Logic, uma plataforma em nuvem para análise de dados de máquinas. Os desenvolvedores que já estiverem rodando um software (como um app ou cluster em nuvem) podem usar a Sumo Logic como backend de seus logs de sistemas. A Sumo Logic proporciona inteligência contínua para dados de máquina.

24

Analisando e prevenindo o preconceito inconsciente em Machine Learning

Este artigo é baseado na palestra de Rachel Thomas, "Analisando e Prevenindo o preconceito inconsciente na Aprendizagem de Máquina" apresentado na QCon.ai 2018. Thomas trabalha na fast, um laboratório de pesquisa sem fins lucrativos que possui parceria com o Instituto de Dados da Universidade de São Francisco em fornecer treinamento em aprendizado profundo para a comunidade de desenvolvedores.

30

Podemos confiar em algoritmos para tomada de decisão automática?

A adoção de tomada de decisão automática vem crescendo a cada dia. Os algoritmos podem produzir resultados socialmente não compreendidos. Há como afirmar que são seguros se não podemos compreendê-los? Os receios do público sobre a incapacidade de prever as consequências adversas impediram tecnologias como a energia nuclear e as culturas geneticamente modificadas.

NOS ACOMPANHE



/InfoQBrasil
/qconsp/



@InfoQBrasil
@QConSP



/company/qcon-
são-paulo/

FALE CONOSCO

FEEDBACK feedback-br@infoq.com

VENDAS sales-br@infoq.com

EDITORIAL editor-br@infoq.com

CARTA DO EDITOR



Srini Penchikala

Tecnologias de Machine Learning (ML) e Deep Learning, como o Apache Spark, o Flink, o CNTK da Microsoft, o TensorFlow e o Caffe aproximaram a análise de dados para a comunidade de desenvolvimento. Seja classificando dois milhões de produtos a serem vendidos recebidos de mais de 700 vendedores multinacionais na organização “Love the Sales”, construindo consciência de algum viés oculto com clientes no site Einstein da Salesforce ou entendendo o comportamento de um sistema de software com Machine Learning e dados ordenados cronologicamente na SumoLogic. As soluções que fazem uso de Machine Learning estão guiando a margem competitiva em empresas e indústrias.

Essa eMag foca no cenário atual de tecnologias de Machine Learning e apresenta diversos estudos de caso do mundo real associados ao tema. Ela apresenta artigos e entrevistas cobrindo diversos tópicos, incluindo:

- Usando Algoritmos de Support Vector Machines (SVMs) como uma ferramenta efetiva para classificação de documentos;
- Analisando e prevenindo vieses inconscientes em machine learning;

- Explorando como a cidade de Nova York estabeleceu uma força-tarefa para obter explicação e mitigação de pessoas afetadas pelo uso de algoritmos de machine learning pelas agências da cidade;

Todas as publicações incluídas nesta emag são escritas por pessoas com experiência prática e especialistas nas matérias apresentadas, todos no campo de machine learning. Esperamos que você concorde conosco que estes artigos são recursos valiosos de referência e que as técnicas podem ser utilizadas em seus próprios projetos e iniciativas em suas organizações.

Como disse Einstein, “educação não é aprender os fatos, mas treinar a mente a pensar”. Nós no InfoQ esperamos que essa emag ajude você a se desenvolver com casos do mundo real de como Machine Learning está sendo usada por diferentes companhias e também agir como catalisador na busca por mais e mais inovações e usos de aplicação das técnicas e algoritmos de Machine Learning.

Obrigado por conferir mais essa eMag do InfoQ.

COLABORADORES



Srini Penchikala

Srini Penchikala trabalha atualmente como um arquiteto de software sênior em Austin, Texas. Penchikala tem mais de 22 anos de experiência em arquitetura de software, design e desenvolvimento. Ele também é o líder editorial para a comunidade de [AI, ML e Engenharia de dados](#) do InfoQ, que atualmente publicou seu minibook [Processamento de Big Data com Apache Spark](#). Ele publicou artigos sobre arquitetura de software, segurança, gerenciamento de riscos, NoSQL e big data em sites como o InfoQ, TheServerSide, O'Reilly Network (OnJava), DevX's Java Zone, Java.net, e JavaWorld.



David Bishop

Depois de estudar ciência da computação na Nova Zelândia, David Bishop se mudou para Londres e liderou o time técnico do reed.co.uk, um dos 100 maiores sites de emprego do Reino Unido. Ele fundou seu próprio negócio de tecnologia, o [Love the Sales](#), que procura agregar todas as vendas de milhares de sites de comércio.



Michael Stiefe

Principal na Reliable Software, Inc, é um consultor de arquitetura de software e desenvolvimento, e o alinhamento de tecnologia da informação com metas de negócios. Ele deu aulas no Departamento de Aeronáutica e Astronáutica do Instituto de Tecnologia de Massachusetts, onde sua pesquisa e foco de docência foi em entender como pessoas constroem modelos mentais para resolver problemas. Como professor adjunto, ele ensinou graduandos nos cursos de engenharia de software na Northeastern University e na Framingham State University. Ele explora seu interesse em tecnologia e arte no blog [Art and Software](#).



Mayukh Bhaowal

É diretor de gestão de produtos na Salesforce Einstein, trabalhando em machine learning automatizado e ciência de dados. Mayukh é mestre em ciência da computação pela Universidade de Stanford. Antes da Salesforce, ele trabalhou em startups na área de machine learning e analytics. Ele atuou como chefe de produto na Scaled Interference, uma startup de plataforma de machine learning apoiada pela Khosla Venture e gerenciou produtos na Narvar, uma startup de e-commerce apoiada pela Accel. Ele também foi um gerente de produtos principal no Yahoo e Oracle.



Roland Meertens

É um engenheiro de visão computacional trabalhando com inteligência artificial de percepção em carros autônomos Autonomous Intelligent Driving, uma subsidiária da Audi. Ele trabalhou em coisas interessantes como tradução neural de máquinas, fuga de obstáculos em drones pequenos e um robô social para idosos. Além de escrever notícias sobre machine learning no InfoQ, ele algumas vezes publica em seu blog [Ping of Intelligence](#) e no Twitter.



PONTOS PRINCIPAIS

- As SVM's (máquina de vetores de suporte) são uma ferramenta eficaz para classificar documentos.
- Ao reduzir o tamanho de grandes conjuntos de dados/vetores, o treinamento de seus modelos é facilitado.
- Ao reutilizar dados rotulados por meio de relacionamentos vinculados, o custo do treinamento de cargas de dados é reduzido e aumenta a precisão das previsões.
- Escolher as estruturas de dados corretas é muito importante para alcançar os melhores resultados.
- Diminuir a hierarquia de dados pode ser útil para reduzir o número de SVMs.

MAIS BYTES NO SEU BOLSO

por **David Bishop**

Em muitos casos, a aquisição de dados de treinamento bem rotulados é uma grande dificuldade para desenvolver sistemas de predição acurados com aprendizado supervisionado.

Na [Love the Sales](#), agregamos produtos de venda de mais de 700 fornecedores internacionais, resultando em mais de 2 milhões de produtos por dia que precisam de classificação. Com uma equipe tradicional de merchandising, seriam necessários 4 anos para realizar esta tarefa manualmente.

Nosso desafio foi aplicar a classificação ao metadata textual destes 2 milhões de produtos (a maioria roupas e utensílios domésticos) em mais de 1000 categorias diferentes - representados em uma hierarquia, como esta:

- Mens Clothing
 - Mens Jeans
 - Mens Jumpers
- Womens Clothing
 - Womens Jeans
 - Womens Jumpers
- ...

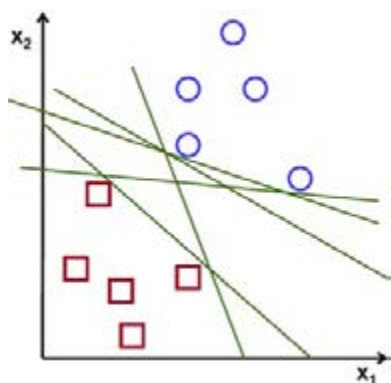
Máquinas de Vetores de Suporte

Para a classificação, optamos pelas SVMs. As SVMs são uma classe de algoritmo de aprendizado de máquina supervisionado, algo apropriado para a classificação de dados linearmente separáveis.

Essencialmente, dado um conjunto suficientemente grande de dados de treinamento rotulados - uma SVM irá tentar encontrar um plano melhor entre os exemplos - ou seja, desenhar uma primeira linha multidimensional para encontrar chão.

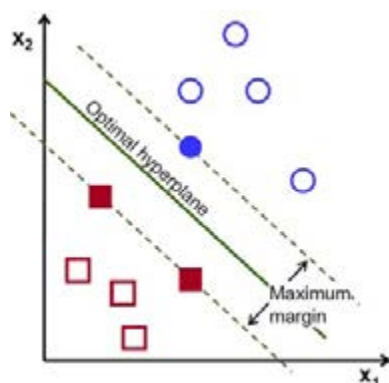
Saiba mais sobre SVM's [aqui](#) e [aqui](#).

Aqui, temos algumas possibilidades de separar este conjunto de dados:



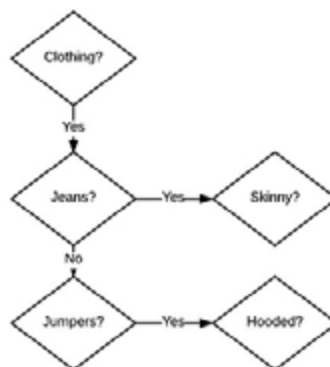
Fonte da imagem: opencv.org.

A SMV tentará aprender o melhor hiperplano:



Embora existam inúmeros algoritmos de machine learning para classificação ([Redes Neurais](#), [Random Forest](#), [Naive Bayesian](#)), as SVMs são ótimas para dados com muitas características - no nosso caso, para classificação de documento, onde cada 'palavra' é tratada como uma variável discreta.

As SVMs podem ser classificadas em múltiplas classes, mas optamos por utilizar uma hierarquia de duas classes simples de SVMs, ligadas de forma hierárquica.



A principal razão para isto é que, quando tentamos, nos pareceu produzir resultados melhores; e, importante ressaltar, utilizou-se muito menos memória em nossa plataforma de aprendizado de máquina, pois cada SVM só necessita saber sobre duas classes de dados. A grande utilização de conjuntos de dados (mais de 300 exemplos) e grandes vetores de entrada (1 milhão de pa-

Essencialmente, dado um conjunto suficientemente grande de dados de treinamento rotulados - uma SVM irá tentar encontrar um plano melhor entre os exemplos - ou seja, desenhar uma primeira linha multidimensional para encontrar chão.

A **Stemização** é uma técnica comum e útil quando se lida com grandes quantidades de textos, com objetivo de escolher palavras diferentes porém com significados e radicais parecidos, e então “reduzir” a um símbolo comum. Por exemplo, as palavras “gato”, “gata”, “gatos”, “gatas” têm significados semelhantes. Quando aplicado o algoritmo “Porter stemmer”, o resultado é “gat”; fazendo desta forma cortamos pela metade o número de palavras para nos preocupar. Usando sistemização em conjuntos com a remoção de palavras “ruidosas” (palavras repetitivas que não trazem significados como “o”, “a”, “os”, “as”, “e”, “com...”) pudemos chegar a um número reduzido e possível de trabalhar.

Uma vez que tenha pré-processado seu conjunto de texto, o próximo passo é treinar seu modelo. Para isto, primeiro é preciso transformar seus textos em um formato que a SVM possa entender - isto é conhecido como “vetorização”. A seguir, temos uma simples descrição do processo para a seguinte sentença:

Após o pré-processamento como descrito acima (sistemização e remoção de palavras):

Utilizando apenas as palavras do exemplo anterior, podemos ver que uma palavra se repete, então podemos transpor o dado da seguinte forma:

Occurrences	Term
1	fantastic
1	great
1	jean
2	men
1	pair
1	skinny

Isso funciona bem para um conjunto de poucos termos. Porém, conforme adicionamos mais e mais exemplos, nosso vocabulário aumenta. Por exemplo, quando adicionamos outro exemplo que não é jeans skinny masculino: *“women bootcut acid wash jean”*.

Isto significa que nosso vetor inicial de termos para jeans skinny masculinos foi alterado para: $[0, 0, 1, 1, 1, 2, 1, 1, 0, 0]$.

Quando lidamos com milhares de fontes, nosso vocabulário começa a ficar grande, tornando-se cada vez mais pesado, e assim os exemplos para treinamento começam a ficar vazios em sua maioria e muito grandes: $[\theta, 0, 0, 0, 0, 0, 0, 0, \dots, 2, 0, 0, 0, 0, 0, \dots, 1, 0, 0, 0, 0, \dots]$.

não-zeros, e a biblioteca (no nosso caso LibSVM) irá magicamente descobrir os locais corretos e preencher os buracos.

Para isto, deve ser passado o vetor de termos e as classes que representam como “Índice do Termo” relativo ao vocabulário inteiro para todos os exemplos de treinamento que desejar utilizar. Por exemplo:

Term Index	Term
0	acid
1	bootcut
2	fantastic
3	great
4	jean
5	men
6	pair
7	skinny
8	wash
9	women

```
Índice do termo #2 : 1
ocorrência
Índice do termo #3 : 1
ocorrência
Índice do termo #4 : 1
ocorrência
Índice do termo #5 : 2
ocorrência
Índice do termo #6 : 1
ocorrência
Índice do termo #7 : 1
ocorrência
```

E que ainda pode ser transformado em algo mais sucinto: $[2:1, 3:1, 4:1, 5:2, 6:1, 7:1]$.

Alexandre Kowalczyk tem uma explicação ótima para a preparação de vocabulário aqui, assim como outros tutoriais ótimos sobre SVM.

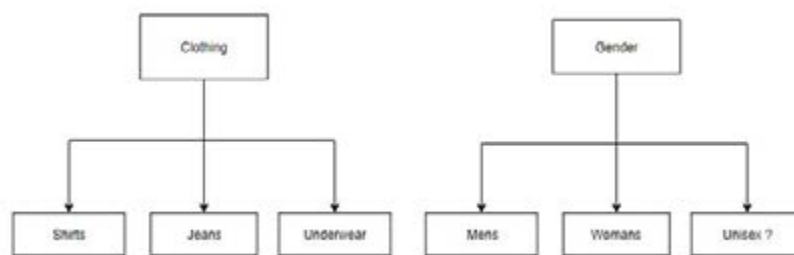
Hierarquia e estrutura de dados

Um aprendizado-chave para nós: a maneira como estas SVMs foram estruturadas pode ter um impacto significativo no quanto de dados de treinamento é necessário aplicar; por exemplo, uma simples abordagem poderia ser como abaixo:



Nesta abordagem, para cada subcategoria nova, duas novas SVM's precisam ser treinadas - por exemplo, a criação de uma nova classe de "Swimwear" iria precisar de uma SVM adicional embaixo de Men's e Women's - sem contar na potencial complexidade de se adicionar a classe "Unisex" no topo. Além disso, grandes hierarquias tornam-se difíceis de se trabalhar.

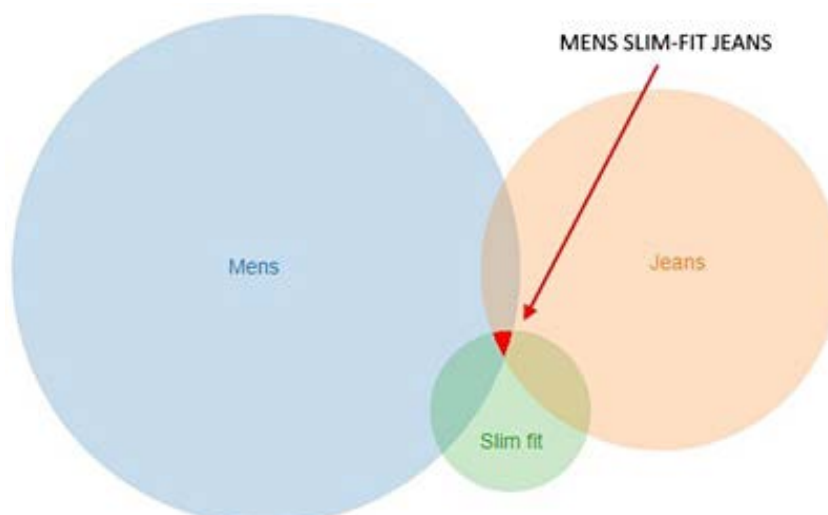
Conseguimos evitar uma grande quantidade de trabalho de rotulagem e treinamento, ao nivelar nossas estruturas de dados em subárvores da seguinte forma:



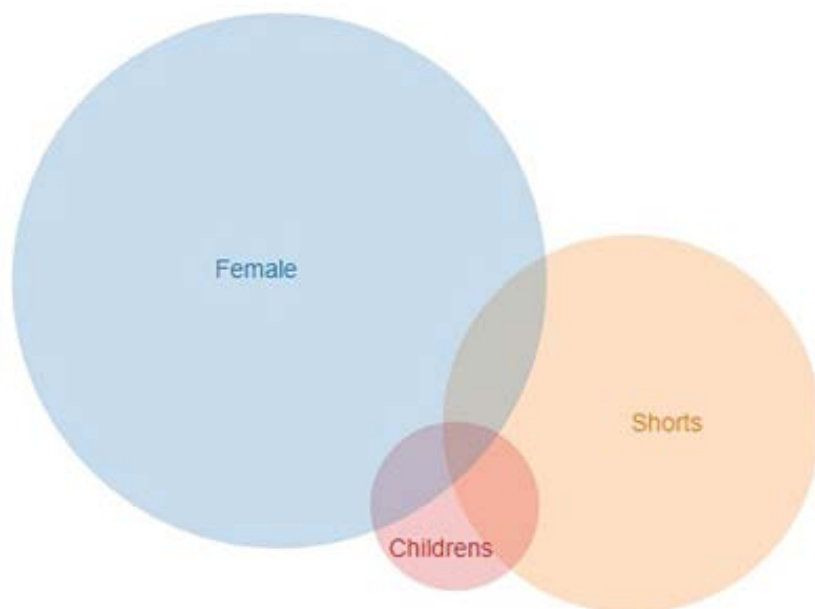
Ao desacoplar nossa estrutura de classificação da hierarquia final, é possível gerar a classificação final percorrendo a hierarquia de SVM com cada documento e verificando os resultados com uma lógica simples baseada em conjuntos, como:

Mens Slim-fit jeans = (Mens and Jeans and Slim Fit) and not Womens

Essa abordagem reduz bastante o número de SVMs necessárias para classificar documentos, pois os conjuntos resultantes podem ser interseccionados para representar a classificação final.



Deve-se notar que ao se adicionar novas classes, abre-se um número exponencialmente crescente de categorias finais. Por exemplo, adicionar uma classe “infantil” no nível superior permitiria imediatamente a criação de uma dimensão inteira de novas categorias infantis (jeans, camisetas, roupas íntimas, etc.), com um mínimos de dados de treinamento adicionais (apenas uma SVM adicional):



Reutilização de dados

Por causa da estrutura que escolhemos, uma das principais informações que conseguimos alavancar foi a reutilização de dados de treinamento, por meio de vinculação de dados relacionados. A vinculação de dados nos permitiu reutilizar nossos dados de treinamento por um fator de 9x - reduzindo assim enormemente o custo e aumentando a precisão das previsões.

Para cada classe individual, obviamente queremos o maior número possível de exemplos de dados de treinamento, cobrindo ambos resultados possíveis. Mesmo que tenhamos construído excelentes ferramentas internas, principalmente uma interface de usuário rápida para pesquisar, classificar e rotular exemplos de dados de treinamento em grandes lotes - rotular milhares de exemplos de cada tipo de produto ainda pode ser trabalhoso, caro e propenso a erros. Determinamos que a melhor maneira de contornar esses problemas era tentar reutilizar todos os dados de treinamento que pudéssemos, em todas as classes.

Por exemplo, considerando algum conhecimento básico de domínio das categorias, sabemos com certeza que “máquinas de lavar roupa” nunca podem ser “limpadores de carpetes.



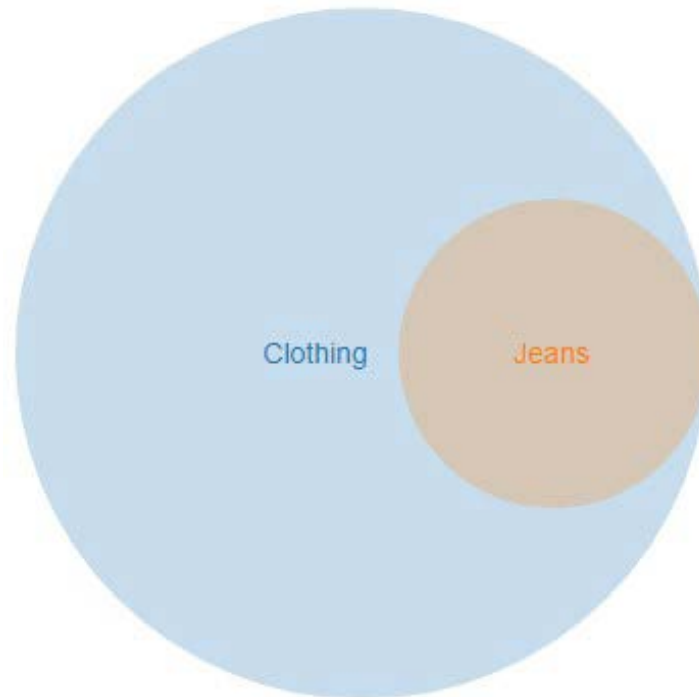
Ao adicionar a capacidade de vincular “Excluir dados”, podemos reforçar a quantidade de exemplos de treinamento “Negativos” para a SVM “máquinas de lavar”, adicionando os dados de treinamento “Positivos” da SVM “limpadores de carpetes”. De maneira mais simples, uma

vez que sabemos que “limpadores de carpetes nunca podem ser “máquinas de lavar roupa” - podemos também reutilizar esses dados de treinamento.

Essa abordagem tem um bom aumento, pois sempre que for necessário adicionar alguns dados de treinamento adicionais para melhorar a SVM “limpadores de carpetes” - ela melhora sem saber a classe “máquinas de lavar”, por meio de dados negativos vinculados.

Por último, outra chance de reutilização (ao considerar uma hierarquia), são os dados de treinamento positivos para qualquer nó filho, pois também são sempre dados de treinamento positivos para seu pai.

Por exemplo: “Jeans” são sempre “Roupas”.



Isso significa que, para cada exemplo positivo de dados de treinamento adicionados à SVM “Jeans”, um exemplo positivo adicional também é adicionado à SVM “Vestuário” por meio de uma vinculação.

Adicionar dados vinculados é muito mais eficiente do que rotular manualmente milhares de exemplos.


Adding training data to: Washing machines

Data 'is': 348 • linked 654 Data 'is not': 259 • linked 13,087

Conclusão

Cremos que as Máquinas de Vetores de Suporte nos ajudaram a alcançar uma qualidade e velocidade de classificação que nunca poderíamos alcançar sem aprendizado de máquina. Como tal, aprendemos que as SVMs são um excelente complemento para qualquer toolkit de desenvolvedores, e que qualquer investigação também deve servir como uma boa introdução a alguns conceitos-chave de aprendizado de máquina.

Além disso, quando se trata das especificidades dos sistemas de classificação hierárquica, desacoplar o componente de classificação da hierarquia resultante, nivelar a estrutura de dados e possibilitar a reutilização dos dados de treinamento será benéfico para obter o máximo de eficiência possível. As abordagens descritas acima não só ajudaram a reduzir a quantidade de dados de treinamento que precisávamos rotular, mas também nos deu uma maior flexibilidade geral.



Essencialmente, dado um conjunto suficientemente grande de dados de treinamento rotulados - uma SVM irá tentar encontrar um plano melhor entre os exemplos - ou seja, desenhar uma primeira linha multidimensional para encontrar chão.



PONTOS PRINCIPAIS

- O viés nos dados criou um gargalo na IA corporativa que não pode ser resolvido por meio da otimização excessiva de algoritmos de aprendizado de máquina ou pela invenção de novos algoritmos;
- O viés cognitivo é a presença accidental de informações nos dados de treinamento que nunca estarão legitimamente disponíveis em produção. Em termos leigos, é como o Marty McFly (em De volta para o futuro) viajando para o futuro, colocando as mãos no Almanaque Esportivo e usando-o para apostar nos jogos do presente;
- Não há bala de prata que resolva isso. Uma combinação de métodos estatísticos e recursos de engenharia podem ajudar a detectar e corrigir este efeito;
- Recursos que exibem esse viés precisam ser diferenciados dos preditores verdadeiros e com isto, determinar o limite correto é a chave fundamental;
- No Salesforce Einstein, a conscientização sobre esse viés com nossos clientes foi o primeiro obstáculo, antes que pudéssemos resolvê-lo

DE VOLTA PARA O FUTURO: DESMISTIFICANDO O VIÉS COGNITIVO

por: **Mayukh Bhaowal**

Era uma vez, um executivo que acompanhava os leads de vendas informando os dados mínimos necessários para inserir um registro de lead. A entrada de dados é uma dor, todos sabemos disso! Enquanto ele trabalhava no processo de conversão dos leads, alguns deles se transformavam em compras. No momento da conversão, ele preenchia informações adicionais apenas para aqueles que tinham o resultado positivo de conversão em compras.

Se treinar seu algoritmo de aprendizado de máquina com anos de tais dados rotulados, ele correlacionará esses recursos com um rótulo positivo, embora eles nunca estivessem realmente disponíveis antes da conversão. O processo de negócios criou um viés nos dados desde o início..

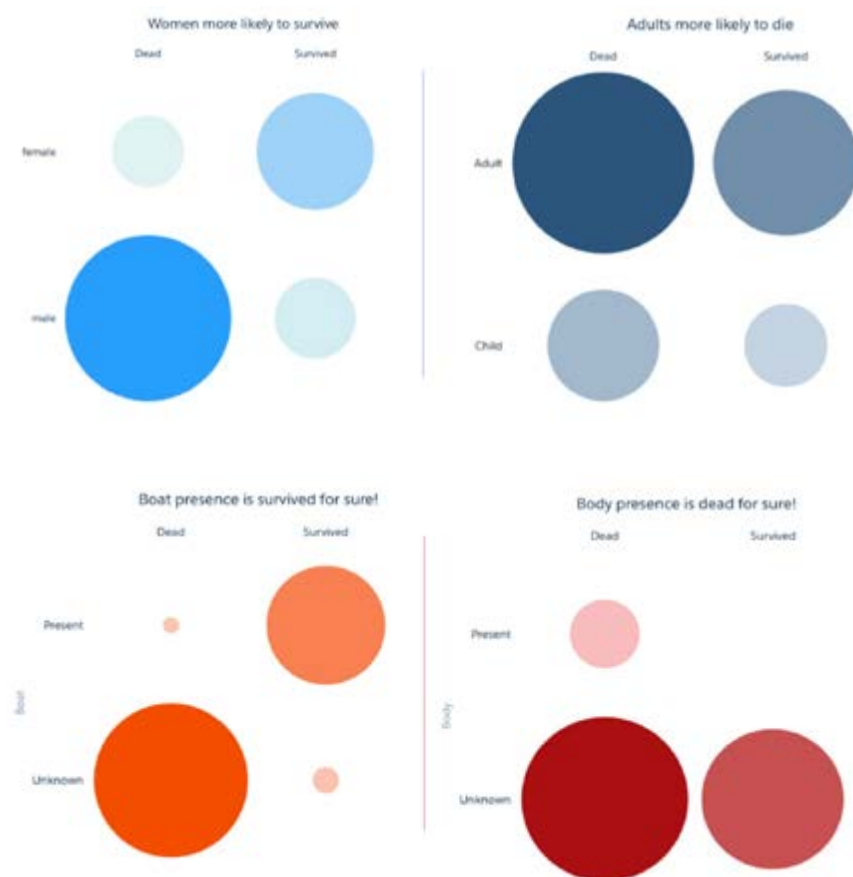
Essa história se repete em diferentes casos de uso, usuários e dados corporativos. Algoritmos de aprendizado de máquina frequentemente assumem que um mítico “conjunto de dados perfeito” é alimentado para prever a rotulação desejada. Na realidade, muitas vezes há muito ruído nos dados. O calcanhar de Aquiles neste domínio é o Hindsight Bias (também conhecido como label leakage ou data leakage). É a presença acidental de informações nos dados utilizados para treinamento que nunca estarão legitimamente disponíveis em um ambiente de produção, causando resultados irreais no ambiente de pesquisa, levando a resultados ruins no ambiente de produção.

Albert Einstein certa vez descreveu o seguinte cenário: “Se tivesse uma hora para resolver um problema, dispensaria 55 minutos pensando no problema e 5 minutos pensando em soluções.

Então, vamos nos aprofundar neste problema um pouco mais, com um exemplo:

Desmistificando o viés cognitivo utilizando o Titanic

Na comunidade de aprendizado de máquina, a previsão de sobrevivência do **Titanic** é bem conhecida. A falta de salva-vidas suficientes foi responsável por muitas vidas perdidas após o naufrágio. Grupos específicos de passageiros, como mul-



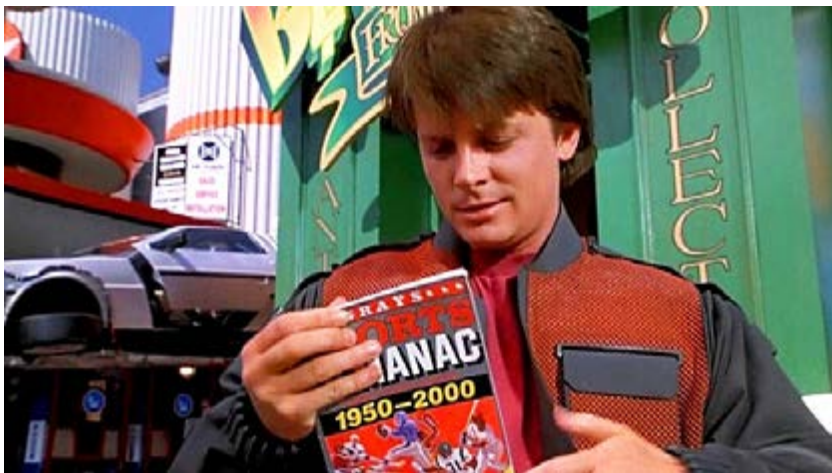
heres, crianças e a classe alta, tinham maior probabilidade de sobreviver do que outros. O aprendizado de máquina é usado para identificar esses sinais e prever quais passageiros sobreviveriam à tragédia.

O que muitos não sabem é que os dados utilizados no [desafio de Kaggle](#) trata-se da versão filtrada e limpa. Os [dados originais](#) possuíam recursos adicionais, dois dos quais eram particularmente problemáticos: os campos Boat e Body. No rescaldo do naufrágio, era atribuído aos passageiros um número de barco, caso chegassem em segurança a um barco salva-vidas, ou um número de corpo, caso fossem eventualmente encontrados mortos. Bem, claro! Se houver um número de corpo, o passageiro está morto. Você não precisa de um algoritmo sofisticado de aprendizado de máquina para lhe dizer isso.

Ao utilizar o conjunto de dados original, as informações sobre o rótulo desejado foram inseridas nos dados de treinamento. Barco e corpo só são conhecidos no futuro após o evento já ter ocorrido. Eles

Cabin class B/C more likely to survive





não são conhecidos no presente ao fazer a previsão. Se treinarmos o modelo com esses dados, ele terá um desempenho ruim no presente, já que essa informação não estaria legitimamente disponível.

Este problema é conhecido formalmente como viés cognitivo. E ocorre predominante em dados do mundo real, que testemunhamos em primeira mão ao criar aplicações preditivas no Salesforce Einstein. Aqui está um exemplo real no contexto da previsão da conversão do lead de vendas: os dados tinham um campo chamado deal value, que era preenchido intermitentemente quando um lead era convertido ou estava próximo de ser convertido (semelhante aos campos Boat e Body na história do Titanic).

Em termos leigos, é como o Marty McFly (em De volta para o futuro) viajando para o futuro, colocando as mãos no Almanaque Esportivo e usando-o para apostar nos jogos do presente. Como a viagem no tempo ainda está a alguns anos, o viés Cognitivo é um problema sério hoje em dia.

O viés cognitivo versus a modelagem de algoritmo

Algoritmos de Aprendizado de Máquina ocupam o centro do

palco hoje em aplicações de inteligência artificial. Há uma corrida para ganhar uma fração de uma melhoria percentual na precisão do modelo, otimizando os algoritmos de modelagem ou inventando novos. Embora isso seja útil, é possível obter um retorno maior para o investimento, focando onde o gargalo é o aprendizado de máquina aplicado, especificamente com dados corporativos. O viés cognitivo é uma dessas áreas, em sua maioria inexplorada. Então, como podemos resolver esse problema?

Estratégias para mitigação

1. Análise estatística para recursos de entrada

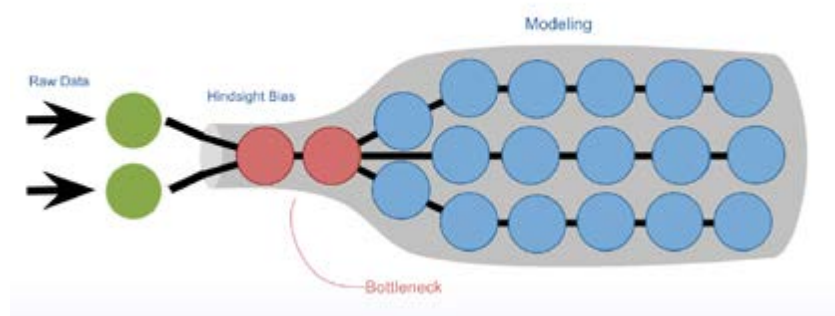
Há um conjunto de testes estatísticos que podemos executar nos recursos de entrada para detectar uma forte associação dos recursos ao

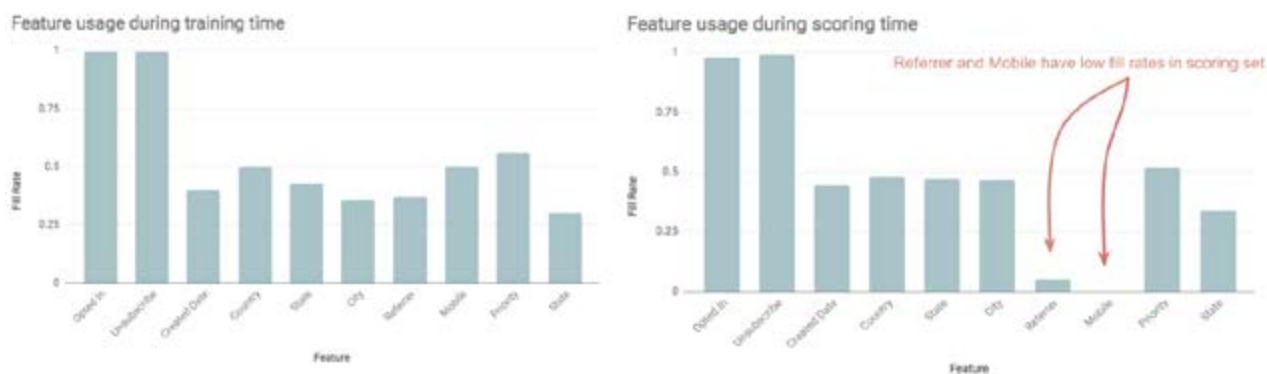
rótulo desejado. A Correlação de Pearson fornece uma medida numérica no intervalo $(-1,1)$ entre o recurso e o rótulo, que expressa a intensidade da associação entre o recurso e o rótulo, bem como a direção. Embora funcione muito bem para recursos numéricos, ele também pode funcionar para recursos categóricos assim que forem vetorizados. No entanto, se os categóricos tiverem um grande número de valores exclusivos (por exemplo, cidades no mundo), a correlação perderá a associação com rótulos devido à diluição do recurso em várias colunas durante a vetorização. Isso pode ser resolvido utilizando CramersV e, portanto, é um teste estatístico mais preferido para recursos categóricos.

O impacto de tais características tendenciosas pode ser mais complicado quando afeta uma pequena fração dos exemplos. Imagine dados geográficos globais. A parte das linhas em que City = San Francisco pode ser uma em mil. O Lift é uma medida alternativa que captura essa dispersão do Viés Cognitivo.

2. Análise estatística para recursos derivados

Uma estratégia que se mostrou útil é executar alguma engenharia preliminar de recursos antes de





executar testes estatísticos nos recursos de entrada.

Por exemplo, muitas características categóricas com viés cognitivo seguem o padrão de ser nulo, até que o rótulo desejado seja determinado. Eles tendem a ter algum valor preenchido, próximo ao quanto o rótulo é especificado. Os campos boat e body, por exemplo, dos dados do Titanic são exemplos desse padrão. A maneira de eliminá-los é adicionar um recurso derivado do indicador nulo (isNull) e usar o CramersV como um teste estatístico.

A correlação nem sempre captura recursos numéricos com viés cognitivo. Por exemplo, no contexto de prever se em uma oportunidade de vendas haverá ganho ou perda, havia um recurso chamado receita esperada. O sistema preencheu o valor depois que o vendedor fechou a oportunidade. Quando o vendedor perdeu a oportunidade, o sistema calculou a receita esperada como 0 ou 1. Caso contrário, o sistema a calculou como um número grande. Uma árvore de decisão pode ser usada para descobrir as duas faixas: [0,1] e [2, infinito]. Depois de colocar um recurso numérico, é possível tratá-lo como um recurso categórico.

Um teste estatístico como o CramersV pode então revelar a forte associação entre o bin específico e o rótulo, expondo assim o viés.

O outro padrão digno de nota que observamos: características categóricas disfarçadas de texto. Por exemplo, ao prever se em um acordo haverá perda ou ganho, havia um recurso chamado Lost no palco. Claramente, fortemente tendencioso, foi definido como um recurso de texto, mas com apenas três valores possíveis. Uma checagem de cardinalidade em tais recursos, convertendo-os em categóricos e, em seguida, aplicando os testes estatísticos de CramersV pode revelar um viés cognitivo.

3. Treinamento versus score de distribuição

Algumas vezes, o viés cognitivo mais indescritível pode não ser exposto às técnicas apresentadas anteriormente apenas olhando para os dados de treinamento. Uma das principais suposições por trás do treinamento de um algoritmo de aprendizado de máquina é que os dados usados para treinamento são semelhantes aos dados utilizados para score.

Como os recursos com viés retrospectivo contêm infor-

mações sobre o rótulo no momento ou logo antes de o rótulo real ser determinado, podemos observar a distribuição dos recursos nos dados de treinamento e os dados de score (antes de conhecer o rótulo real). Se alguma das características apresentar uma lacuna estatisticamente significativa nas duas distribuições, isso é um candidato a viés cognitivo.

O ponto de corte temporal ou por registro de data e hora é uma técnica relacionada. Neste caso, determinamos um timestamp de corte como o momento em que o evento de previsão deve ocorrer, com base nos registros atuais e passados. Em seguida, excluimos todos os dados antes do evento de interesse. Por isso, não usamos nenhum dado que coletamos perto da previsão ou depois, ou seja, no futuro.

4. Validação cruzada e a preparação de dados

É crucial executar toda preparação de dados e engenharia de recursos em cada validação cruzada. Por exemplo, se usarmos as informações do rótulo em qualquer etapa de engenharia de recursos, como a categorização, introduzimos inerentemente o viés cognitivo nos dados. O mesmo se aplica aos métodos de seleção de recursos, remoção de

outliers, codificação e dimensionamento de recursos para redução de dimensionalidade. Se executarmos qualquer um deles nos dados inteiros antes da validação cruzada, então os dados de teste em cada dobra do procedimento de validação cruzada desempenharam um papel na escolha dos recursos, e isso introduz um viés cognitivo nos dados.

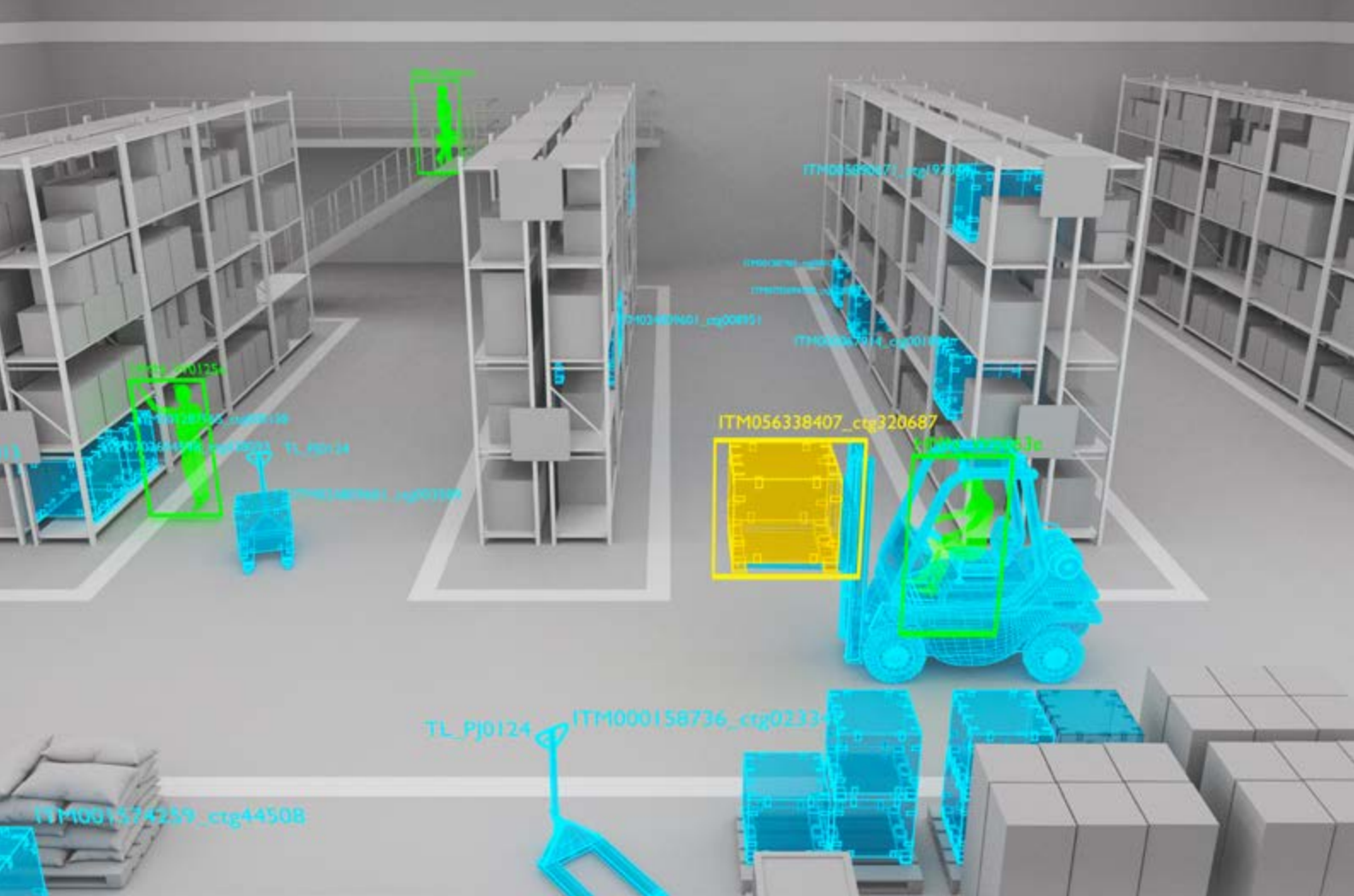
Isto é um viés cognitivo ou uma predição verdadeira?

Em todos os métodos discutidos até agora, o aspecto mais difícil é descobrir o limite certo para os seus dados e caso de uso, o que ajudaria a revelar um viés cognitivo. Qual deve ser a medida de correlação, além da qual um recurso é considerado como tendencioso? 0,9 é um bom limiar ou deveria ser 0,75? Em que ponto um recurso é tendencioso versus realmente um verdadeiro preditor? É preciso tomar a mesma decisão em todas as outras medidas estatísticas, incluindo a diferença na distribuição de treinamento e pontuação e assim por diante.

No Salesforce Einstein, nossa experiência na criação de modelos para uma ampla variedade de casos de uso e dados de diferentes formas e tamanhos ajuda a informar limites aceitáveis. No entanto, está longe de ser cravado na pedra. Estamos continuamente fazendo iterações nos limites para refletir os dados e problemas do mundo real.

Conclusão

O viés cognitivo na IA corporativa é um problema mais prevalente quando comparada à IA na academia ou para um consumidor. O desafio mais significativo que enfrentamos foi a conscientização com nossos clientes. Depois que passamos por isso, entender os processos de negócios e os padrões de dados que introduzem esse viés foi crucial. Essa jornada nos ajudou a desenvolver soluções que automatizam a detecção do viés cognitivo. O resultado que encontramos foram previsões de aprendizado de máquina mais confiáveis.



PONTOS PRINCIPAIS

- Antes de entrar em machine learning para o comportamento sistêmico de softwares, deve-se ter conhecimento sobre os conceitos de séries temporais.
- Dados faltantes na sua série temporal podem levar a resultados inesperados enquanto estiver analisando-os. A “biblioteca Pandas” pode ajudar a trabalhar com o preenchimento destes valores de uma forma sensata.
- Quando humanos estão usando seu serviço, espere pela sazonalidade em seus dados. Leve em conta este detalhe quando for desenhar seus algoritmos preditivos.
- Tome cuidado com o limite definido no momento da detecção de anomalia. Eventos que são improváveis para um simples servidor, tornam-se muito prováveis quando estiver dimensionando sua aplicação.
- Entenda o que estiver tentando alcançar quando estiver analisando séries temporais. Tenha certeza de não usar análises determinísticas como a linguagem SQL permite. Conheça o comportamento de seu algoritmo em uma escala matemática e se realmente está automatizando a interpretação deste, ou se está transformando dados em resíduos preditivos e os usando em suas análises.

ENTENDENDO O COMPORTAMENTO DE SISTEMAS E SOFTWARES COM MACHINE LEARNING E DADOS DE SÉRIES TEMPORAIS

por **Roland Meertens**

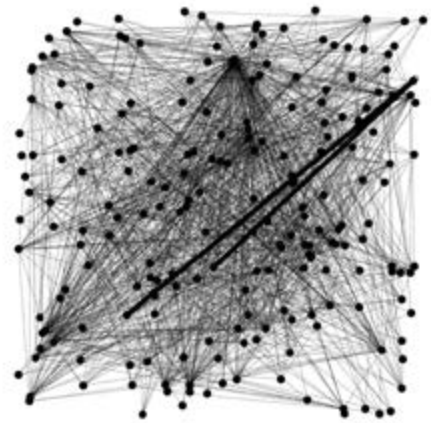
No [QCon.ai 2018](#), David Andrzejewski apresentou “Entendendo o Comportamento Sistêmico de Softwares com Machine Learning e dados de séries temporais”. David é gerente de engenharia na Sumo Logic, uma plataforma em nuvem para análise de dados de máquinas. Os desenvolvedores que já estiverem rodando um software (como um app ou cluster em nuvem) podem usar a Sumo Logic como backend de seus logs de sistemas. A Sumo Logic proporciona inteligência contínua para dados de máquina.

Muitas coisas rodam em softwares, e técnicas de inteligência artificial estão entrando no mundo de softwares. Antes de entrar a fundo no impacto que o machine learning proporciona no comportamento sistêmico de softwares, é preciso entender as abordagens tradicionais relacionadas às séries temporais. Conhecer as limitações dos métodos tradicionais permite que faça trocas conscientes ao optar por alguma técnica. Primeiro, pergunte a você mesmo se conhece o que está tentando realizar. Uma vez que saiba, se questione se é possível cumprir isto com uma análise simples ou determinística. Olhe apenas para o machine learning quando os outros métodos forem impossíveis.

Entender o que seu software está fazendo e o porquê de estar falhando pode ser difícil. As empresas que implantam serviços que dependem de muitos outros microserviços em vários servidores podem se beneficiar de um diagrama que lista as dependências entre estes microserviços. Ao desenhá-lo, pode-se ter uma imagem do que as pessoas chamam de “estrela da morte” de microserviços:

Muitas aplicações geram terabytes de logs por dia, que consistem de gigabytes de código fonte e geram milhões de métricas por minuto. Analisar este

Microservices “death star”



dado manualmente é impossível, então é preciso a inteligência de máquina. Entretanto, analisar os dados e encontrar apenas o que o seu sistema estiver REALMENTE fazendo é uma tarefa difícil senão impossível. Um artigo que vai mais a fundo na granularidade do dado e em qual momento você precisa dele é o “[Poderia um neurocientista entender um microprocessador?](#)”. Os autores deste artigo usam um simulador para jogar uma versão antiga de Donkey Kong. Por possuírem a memória da simulação, tiveram acesso ao estado completo do sistema. Teoricamente, isto significa que é possível analisar o dado e tentar fazer uma engenharia reversa no que estiver acontecendo com um nível maior de entendimento, apenas por olhar o dado em detalhe. Embora essa tática possa proporcionar insights pequenos, é improvável que ape-

nas olhar os dados permitirá você a entender completamente um nível maior de Donkey Kong.

Esta analogia torna-se importante para quando estiver usando apenas dados brutos para entender sistemas multiescala, dinâmicos e complexos. Agregar os dados brutos em visões de séries temporais torna o problema mais acessível. Uma boa fonte sobre isso é o livro “Site Reliability Engineering”, que pode ser [lido gratuitamente](#).

Entender sistemas multiescalas, dinâmicos e complexos é especialmente importante para um engenheiro em serviço. Quando um sistema cai, é preciso descobrir o que o sistema está realmente fazendo naquele momento. Por este motivo, o engenheiro precisa dos dados brutos e dos meios para visualizá-los, assim como métricas de alto nível que conseguem sumarizar os dados. Um engenheiro nesta situação normalmente quer entender como este servidor está se comportando quando comparado a outro servidor, ou a ele mesmo no dia anterior, ou a ele mesmo antes de uma atualização do software.

Vantagens e desvantagens dos percentis

Quando olhamos um longo histórico de dados (log), não entramos nos detalhes de milissegundos contínuos. Seus

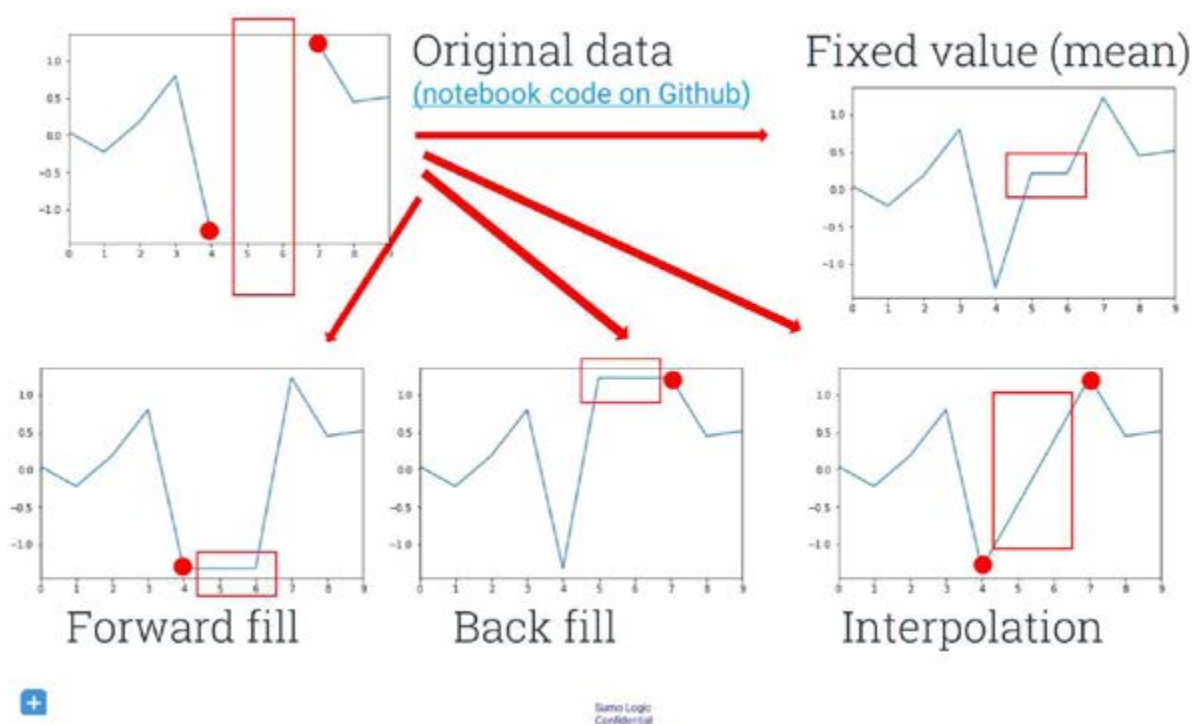
Operational time series telemetry: why

Q: WTF is my system actually doing?

Monitoring & troubleshooting

- data visualization
- alerting*
- summarize behavior
- comparisons





dados são quantificados em tempo. O caminho mais básico para fazer isso é utilizar funções como min, max, average (média), sum (soma) e count (contagem). Muitas pessoas que agregam dados gostam de usar percentis também. A vantagem dos percentis é que podem expressar seus dados em uma linguagem não ambígua. Um exemplo de sentença sem percentil é "O tempo máximo para carregar uma solicitação foi 4.300 milissegundos." Esta sentença é precisa mas não ajuda a determinar quão distante está dos padrões de uma operação normal que falhará. Porém, diga-se que "p99 é menos do que 2.000 milissegundos" indica que não mais que 1% das solicitações de clientes levam mais do que dois segundos para carregar.

A desvantagem dos percentis é que dificultam a combinação de dados em algo significativo. Embora os valores em torno do 50º percentil tendam a ser estáveis, os percentis mais altos variarão muito e têm uma distribuição longa de valores possíveis. Outro

problema é ser fácil agregar as análises simples de vários conjuntos de dados. Pode-se calcular o mínimo de dois conjuntos de dados observando apenas os mínimos de ambos. No entanto, não se pode simplesmente usar os métodos com percentis. É matematicamente impossível combinar a p95 do dataset X e a p95 do dataset Y. Isso significa que é difícil dizer algo significativo sobre uma combinação de vários conjuntos de dados sem muito trabalho.

Conceitos importantes de séries temporais

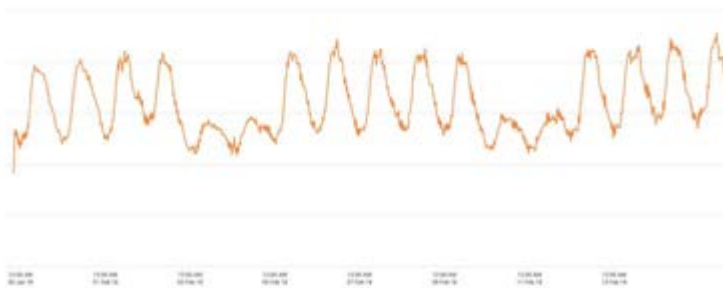
Um aspecto básico de monitoramento para séries temporais são as comparações com mudança de periodicidade. Isso é particularmente importante se quiser comparar a latência de escrita de um cluster com a latência de escrita do mesmo host no dia anterior. Isso também pode ser combinado com "windowing data" (dados de janela), conhecido como "agrupamento ao longo do tempo". Mais informações podem ser encontradas na [palestra do Tyler Akidau durante o QCon](#)

[San Francisco 2016](#), na qual esse conceito foi discutido no contexto do Apache Beam.

O manuseio de dados faltantes também é importante. Antes de aplicar qualquer machine learning, é preciso saber como deseja lidar com os valores ausentes. Colocar valores constantes, como zeros ou infinitos, no lugar de valores ausentes provavelmente levará a resultados inesperados. No entanto, não colocar nada lá provavelmente fará com que tenha exceções de runtime posteriormente no loop. Isso pode ser prevenido usando pandas, uma biblioteca Python de análise de dados, que é um verdadeiro canivete suíço para manipulação de dados. Pode ser usado o método fillna(), que possui alguns valores padrão que são realísticos e sensatos. Observe que há muitas maneiras interessantes de preencher lacunas em seus dados e há muitas formas e métodos que podem ser usados. Algumas áreas chamam isso de "predição" de dados faltantes, outras áreas chamam de "imputação",

Seasonality

Very common in data linked to human activity



“inferência” ou “amostragem”. Você pode usar métodos de preenchimento, simplesmente preenchê-los ou interpolá-los.

Agindo nos dados

A simple thing to think about Uma coisa simples de se pensar ao configurar um sistema de logs é o alerta de limite fixo. O objetivo dos alertas é alertar alguém quando o site cair ou outro evento inesperado. Muitas pessoas iniciam o desenvolvimento de alertas contratando um especialista que pode definir limites sensatos para vários aspectos do sistema. Por exemplo, poderia ser definido um alerta para disparar assim que 5% das solicitações demorem mais de dois segundos, notificando o engenheiro que está em serviço naquele momento.

Contudo, os especialistas humanos não escalam bem. Talvez queira automaticamente comparar o comportamento de algumas máquinas com as de outras máquinas, especialmente quando se tem muitas máquinas disponibilizando muitas séries temporais. Não se pode analisar e comparar todas essas séries temporais sozinho, e um grande número de máquinas pode impedir a comparação entre séries temporais. Este é o ponto onde pode-se tentar aplicar o machine learning.

Modelos preditivos e outliers

Uma abordagem possível é a detecção de outliers usando modelagem preditiva. Ao prever o comportamento normal de suas máquinas, também pode-se detectar quando suas máquinas agem fora da saída esperada. No entanto, é preciso levar muito em consideração antes de se fazer isso. Existem quatro perguntas-chave a se fazer:

- O comportamento é realmente regular?
- Como o comportamento pode ser modelado?
- Como pode ser definido um grande desvio do que é esperado?
- É realmente valioso detectar surpresas e desvios do que é esperado?

Metric similarity: naïve approach

- Are these “behaving similarly”?
- Direct norm distance calculation
 - $d(x, y) = \|x - y\|_2$
 - Spikes are “disjoint”
 - Distance would be large
- Intuition: can we slightly shift?
 - Would be very similar...



Algo importante a se considerar ao fazer a modelagem preditiva é a sazonalidade ou o ritmo de seus dados. Qualquer serviço que tenha humanos no circuito tem potencial para um ritmo. Por exemplo, a maioria das pessoas usa a Sumo Logic no trabalho, o que significa que os dados de uso da Sumo Logic para qualquer país mostrarão muita atividade durante o horário normal de trabalho, mas não tanto fora desse horário. Porém, os dados de uso do Netflix provavelmente mostram uma tendência inversa. Isso pode ser modelado ajustando manualmente seus dados ou usando [transformadas de Fourier](#). Outra opção que muitas pessoas usam são os [modelos ocultos de Markov](#).

Mineração de dados de séries temporais baseada em distância

Quando se tem várias máquinas, provavelmente é desejável comparar o comportamento das máquinas entre si. Se nota-se um comportamento estranho em uma máquina, é desejável descobrir se outras máquinas estão se comportando da mesma maneira. Talvez cada uma esteja executando versões diferentes de software, talvez estejam no mesmo data center ou talvez alguma outra coisa esteja acontecendo. Para analisar isso, deve-se comparar a distância entre as séries temporais.

Qual métrica deve ser usada para determinar a similaridade entre duas séries temporais? Simplesmente diferenciá-las de hora em hora subtraindo uma da outra obrigatoriamente dará resultados errados. Na imagem acima, embora as séries temporais sejam bastante semelhantes, essa métrica dirá que são completamente diferentes.

Existe todo um universo de métricas que pode ser usado. Uma técnica popular é a distorção dinâmica do tempo, que basicamente questiona como se pode transformar, deformar ou distorcer sua série temporal para colocá-los no melhor alinhamento e qual penalidade terá que ser paga por essa modificação. Com essa métrica, pode-se localizar os N hosts que se comportam de maneira mais semelhante ou pode-se criar um gráfico de similaridade de host. O uso de clustering espectral pode fornecer uma imagem que informa sobre qualquer estrutura em seus hosts.

Detecção de anomalias e classificação de eventos com dados de log

Existem maneiras de transformar seus dados de log em uma série temporal. Quando se tem um alto volume de strings semi-estruturadas, pode-se contar as mensagens ou extrair in-

formações delas. Esses logs são um rastreamento aproximado da execução de programa. Como não se pode inserir um depurador para suas máquinas depois de estarem em produção, só é possível deduzir o comportamento do seu software por meio dessas mensagens de log. Se o seu programa imprimir uma string toda vez que uma solicitação expirar, será possível contar o número de tempos limite a cada hora. Isso resultará em uma série temporal, que você acabou de aprender a analisar!

Talvez seja tentado a definir um limite nos valores de certas séries temporais. No entanto, não queira se enganar pensando que encontrou um evento interessante quando, na verdade, o evento não tinha nada demais. Imagine que tenha um modelo super-preciso e deseje enviar um alerta sempre que houver apenas 0,01% de chance de ocorrer um padrão. Com um serviço com um milhão de séries temporais, pode-se esperar cerca de cem falsos positivos. Baron Schwartz, em sua palestra [“Por que ninguém se importa com sua detecção de anomalias”](#), entra em mais detalhes sobre quais técnicas deveriam ser usadas para determinar um limite.

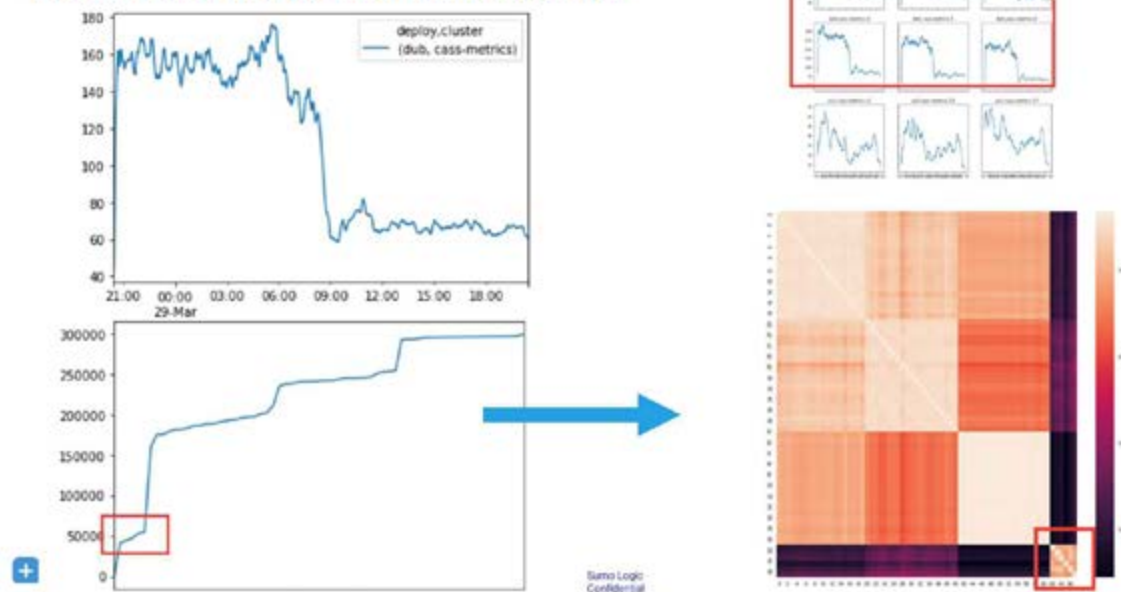
Com todos os avanços recentes em deep learning, talvez queira

usar isso para ajudar nas previsões e detecção de anomalias, mas o deep learning ainda não consegue nos livrar da compreensão do domínio do problema. Ainda é preciso encontrar uma maneira de enquadrar seus problemas. Uma abordagem possível é o uso de redes neurais recorrentes para prever. Essa é uma ótima ideia se tiver acesso a muitos dados de treinamento. Se não, sua primeira prioridade deveria ser a agregação dos dados antes de tentar fazer algo com eles.

Para concluir, a toca do coelho no quesito de dados de inspeção é muito funda. Temos máquinas controlando nossas vidas. E essas máquinas produzem dados, mas analisar os dados é complicado, por isso temos as ferramentas de machine learning, que são delicadas. É de grande importância evitar ruídos e falsos positivos, e para fazer isso, é preciso ter certeza de que se entende o que se está tentando fazer. Saiba por que não está usando análises determinísticas semelhantes ao SQL e entenda os métodos usados em escala matemática. Por fim, saiba se está automatizando a interpretação ou transformando dados em resíduos preditivos e usando isso para previsão de anomalias.

Spectral clustering

[Tutorial \(von Luxborg\), sklearn implementation](#)



Algo importante a se considerar ao fazer a modelagem preditiva é a sazonalidade ou o ritmo de seus dados. Qualquer serviço que tenha humanos no circuito tem potencial para um ritmo... Outra opção que muitas pessoas usam são os modelos ocultos de Markov.



ANALISANDO E PREVENINDO O PRECONCEITO INCONSCIENTE EM MACHINE LEARNING

por **Srini Penchikala**

Este artigo é baseado na palestra de Rachel Thomas, “[Analisando e Prevenindo o preconceito inconsciente na Aprendizagem de Máquina](#)” apresentado na [QCon.ai 2018](#).

Thomas trabalha na fast, um laboratório de pesquisa sem fins lucrativos que possui parceria com o Instituto de Dados da Universidade de São Francisco em fornecer treinamento em aprendizado profundo para a comunidade de desenvolvedores. O laboratório oferece um curso gratuito chamado “Prática em Aprendizado Profundo para Programadores”.

Thomas discutiu o preconceito no aprendizado de máquina, suas fontes e como evitá-los em três estudos de caso.

Estudo de caso 1: Software para sistemas de contratação, demissão e justiça criminal

Algoritmos de aprendizagem profunda estão sendo cada vez mais usados para tomar decisões impactantes, como na

contratação e demissão de funcionários e no sistema de justiça criminal. O preconceito na codificação traz armadilhas e riscos para o processo de tomada de decisão.

O Pro Publica em 2016 [investigou](#) o [algoritmo de reincidência COMPAS](#) que é usado para prever a probabilidade de um preso ou criminoso acusado cometer novos crimes caso liberado. O algoritmo é usado para conceder

fiança, sentenciar e determinar a liberdade condicional. O Pro Publica descobriu que a taxa de falsos positivos (rotulado como “alto risco”, mas não reincidente) foi quase duas vezes maior para réus negros (taxa de erro de 45%) do que para réus brancos (24%).

A etnia não era uma variável explícita inserida nesse algoritmo, mas etnia e gênero são codificados latentemente em muitas outras variáveis, como onde moramos, nossas redes sociais e nossa educação. Mesmo em um esforço consciente para não levar em consideração a etnia ou gênero, não garante a falta de preconceito - supondo que a deficiência visual não funcione. Apesar das dúvidas sobre a precisão do COMPAS, o Supremo Tribunal de Wisconsin confirmou seu uso no ano passado. Thomas argumentou que é horrível que ainda esteja em uso.

É importante ter uma boa base para saber quando um desempenho é bom e ajudar a indicar um modelo mais simples que pode ser mais eficiente. Só porque algo é complicado não significa que funcione. O uso de inteligência artificial (IA) para o policiamento preditivo é uma preocupação.

A Taser adquiriu duas empresas de IA no ano passado e está oferecendo um software preditivo para departamentos de polícia. A empresa detém 80% do mercado das câmeras corporais utilizadas por policiais nos EUA, então eles têm muitos dados de vídeo. Além disso, o Verge [revelou em fevereiro](#) que a polícia de Nova Orleans tem utilizado o software de policiamento preditivo da Palantir nos últimos seis anos em um programa altamente secreto que até mesmo os membros do conselho da cidade não sabiam. Aplicativos como

esses são preocupantes porque não há transparência. Por serem empresas privadas, não estão sujeitas às leis estaduais/públicas da mesma maneira que os departamentos de polícia. Muitas vezes, eles são protegidos no tribunal por terem que revelar o que estão fazendo.

Além disso, há muitos preconceitos raciais nos dados policiais existentes, de modo que os conjuntos de dados dos quais esses algoritmos aprenderão serão tendenciosos desde o início.

Finalmente, repetidas falhas da visão computacional ocorreram ao trabalhar com pessoas negras. Thomas disse que esta é uma combinação assustadora de coisas para dar errado.

Estudo de caso 2: visão computacional

A visão computacional costuma ser ruim para reconhecer pessoas negras. Um dos exemplos mais infames vem de 2015. O Google Fotos, que classifica automaticamente as fotos, classificou fotos de formaturas e imagens de edifícios de maneira útil. Ele

também rotulou pessoas negras como gorilas.

Em 2016, o site [Beauty.AI](#) que usava robôs com IA como juizes em concursos de beleza, descobriu que pessoas com pele clara eram julgadas muito mais atraentes do que pessoas com pele escura. E em 2017, o [Face-App](#), que usa redes neurais para criar filtros para fotografias, criou um filtro de gostosura que iluminou a pele das pessoas e deu-lhes mais recursos europeus. Rachel mostrou um tweet do rosto real de um usuário e uma versão mais sexy dele que o aplicativo criou.

Thomas falou sobre um [trabalho de pesquisa](#) de Joy Buolamwini e Timnit Gebru, que avaliaram vários classificadores comerciais de visão computacional da Microsoft, IBM e Face++ (uma empresa chinesa). Eles descobriram que os classificadores trabalham melhor em homens do que em mulheres, e melhor em pessoas com pele clara do que pessoas com pele escura. Há uma lacuna muito perceptível: a taxa de erro para homens de pele clara é essencialmente 0%, mas varia entre 20% e 35% para



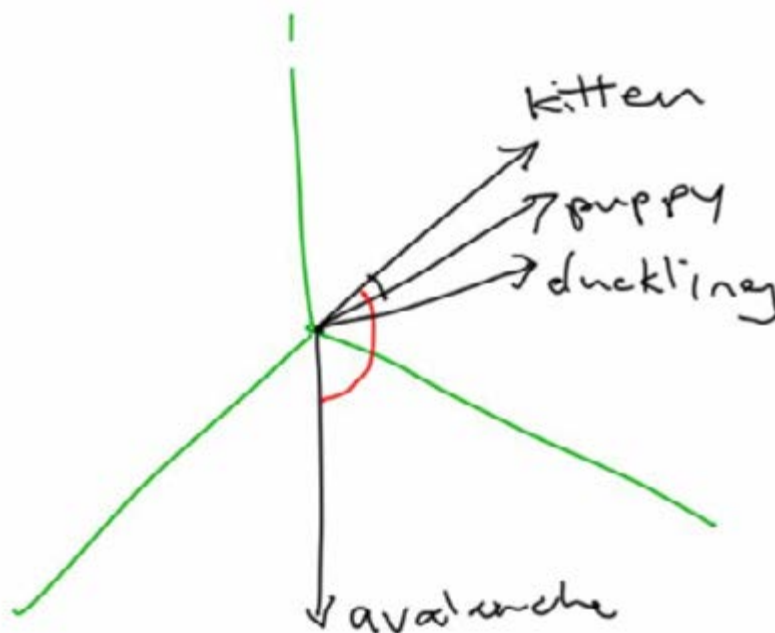
as mulheres de pele escura. Tanto Buolamwini como Gebru também analisaram as taxas de erro para as mulheres por tom de pele. Erros aumentaram com a escuridão da pele. A categoria da pele mais escura apresentava taxas de erro de 25% e 47%.

Estudo de caso 3: incorporando palavras

O terceiro estudo de caso de Thomas é a incorporação de palavras em produtos como o Google Tradutor.

Considere um conjunto de frases como “Ela é médica. Ele é enfermeiro.” Então use o Google Tradutor para traduzi-los para o Turco e depois traduzi-los de volta para o Inglês. Os gêneros misturados e as frases passam a dizer: “Ele é médico. Ela é uma enfermeira.”. O Turco tem um pronome singular neutro de gênero que se traduz em um estereótipo no Inglês. Isso acontece com outras linguagens que têm pronomes singulares que são neutros em relação ao gênero. Foi documentado, por várias palavras, que os estereótipos de tradução sustentam que as mulheres são preguiçosas, que as mulheres são infelizes e muitas outras características.

Thomas explicou o motivo de isto acontecer. Computadores e aprendizado de máquina tratam imagens e palavras como números. A mesma abordagem é usada para o reconhecimento de fala e criação de legendas de imagens. A maneira como esses algoritmos funcionam é que eles pegam uma imagem fornecida e emitem algo como “um homem de camisa preta está tocando guitarra” ou “operário de colete laranja está trabalhando na estrada”. O mesmo mecanismo sugere automaticamente respostas a e-mails de produtos como uma Resposta Inteligente



do Google: se alguém perguntar sobre seus planos de férias, a Resposta Inteligente sugere que se diga “Nenhum plano ainda” ou “Acabei de enviá-los para você”.

Thomas falou sobre um exemplo no curso da [fast.ai](#): “Prática em Aprendizado Profundo para Codificadores”. Neste exemplo, podemos fornecer palavras e recuperar uma imagem. Dado às palavras “tenca” (um tipo de peixe) e “rede” ele retorna uma imagem de uma tenca em uma rede. Esta abordagem passa por diversas palavras e não dá qualquer noção do significado dessas palavras serem semelhantes. Então, “gato” e “catástrofe” pode ser um número sequencial, mas não há qualquer tipo de relação semântica entre eles.

Uma abordagem melhor é representar as palavras como vetores. Os recursos incorporados nas palavras são representados como vetores de alta dimensão. Ela deu um exemplo de “gatinho”, “filhote” e “patinho”, que podem estar todos próximos uns dos outros no espaço, porque todos são

filhotes de animais. Mas o vetor da “avalanche” pode estar longe, já que não há conexão real entre eles.

Para mais informações sobre vetores de palavras, consulte “[O incrível poder dos vetores de palavras](#)”, de Adrian Colyer.

Word2Vec

Word2Vec é uma biblioteca de recursos incorporados de palavras lançado pelo Google. Existem outras bibliotecas semelhantes, como o fastText do Facebook, e o GloVe do Grupo de Processamento de Linguagem Natural da Universidade de Stanford. É preciso muito tempo, dados e poder computacional para treinar essas bibliotecas, por isso é útil que esses grupos já tenham feito isso antes de lançar suas bibliotecas para uso público. É muito mais fácil de usá-la já que esta é uma versão já treinada. O código para os três projetos está disponível no GitHub, assim como o [workshop de incorporação de palavras](#) do Thomas. É possível executar seu programa

usando o Jupyter Notebook e experimentar palavras diferentes.

Os vetores de palavra para palavras semelhantes como “filhote” e “cachorro” ou “rainha” e “princesa” estão mais próximos no eixo. E, claro, palavras não relacionadas como “celebridade” e “poeira” ou “gatinho” e “avião” estão mais distantes. O programa usa uma semelhança de cosseno, não a distância euclidiana, já que não se deseja usar a distância euclidiana em altas dimensões.

Essa solução pode ser usada para capturar algo sobre idioma. Também é possível encontrar as 10 palavras mais próximas de uma palavra-alvo específica. Por exemplo, se procurar as palavras mais próximas a “nadar”, receberá palavras como: “natação”, “remo”, “mergulho”, “vôlei”, “ginástica” e “piscina”. Analogias de palavras também são úteis. Eles captam coisas como “a Espanha é para Madri, como a Itália é para Roma”. No entanto, há muitas oportunidades de preconceito aqui. Por exemplo, a distância entre “homem” e “gênio” é muito menor que a distância entre “mulher” e “gênio”.

Os pesquisadores estudaram cestas de palavras de forma mais sistemática. Eles pegavam uma cesta ou grupo de palavras, como todas as flores: trevo, papoula, calêndula, íris, etc. Outra cesta eram insetos: gafanhoto, aranha, percevejo, larva, etc. Eles tinham uma cesta de palavras agradáveis (saúde, amor, paz, alegria, etc.) e uma cesta de palavras desagradáveis (abuso, sujeira, assassinato, morte, etc.). Os pesquisadores analisaram as distâncias entre essas diferentes cestas de palavras e descobriram que as flores estavam mais próximas de palavras agradáveis e os insetos estavam mais próximos de palavras desagradáveis.

Tudo isso parece razoável até agora, mas depois os pesquisadores analisaram nomes estereotipicamente de pessoas negras e nomes estereotipicamente de pessoas brancas. Eles descobriram que os nomes dos negros estavam mais perto de palavras desagradáveis e os nomes dos brancos estavam mais perto de palavras agradáveis, o que é um preconceito. Eles encontraram uma série de preconceitos raciais e de gênero entre grupos inteiros de palavras, o que produziu analogias como “pai é médico como mãe é enfermeira”, “homem é programador de computador como mulher é dona de casa”. Estas são todas as analogias encontradas no Word2Vec e no GloVe.

Thomas falou sobre outro exemplo de preconceito em um sistema de resenhas de restaurantes

que classificava os restaurantes mexicanos como inferiores, porque palavras incorporadas para “Mexicano” tinham conotações negativas. Estas incorporações de palavras são treinadas com uma quantidade gigante de textos. Esses textos contêm muitos preconceitos raciais e de gênero, pois a palavra incorporada aprende com estas associações ao mesmo tempo em que aprendem os significados semânticos que queremos que eles saibam.

O aprendizado de máquina pode amplificar o preconceito

O aprendizado de máquina pode realmente ampliar o preconceito. Um exemplo disso é discutido em “Os homens também gostam de fazer compras: Reduzir a amplificação do preconceito de gênero usando restrições de nível de

O aprendizado de máquina pode realmente ampliar o preconceito. Um exemplo disso é discutido em “Os homens também gostam de fazer compras”, que analisou a rotulagem semântica de imagens em um conjunto de dados. Os pesquisadores descobriram que 67% das imagens de pessoas que cozinhavam eram mulheres, mas o algoritmo classificou 84% dos cozinheiros como sendo mulheres.

[corpus](#)”, que analisou a rotulação semântica de imagens em um conjunto de dados. Os pesquisadores descobriram que 67% das imagens de pessoas que cozinhavam eram mulheres, mas o algoritmo encontrou 84% dos cozinheiros como sendo mulheres. Existe o risco de algoritmos de aprendizado de máquina amplificar o que vemos no mundo real.

Thomas mencionou a pesquisa de Zeynep Tufekci, que forneceu descobertas sobre a intersecção entre tecnologia e sociedade. Tufekci twittou que “o número de pessoas que me dizem que a reprodução automática do YouTube termina com vídeos de supremacia branca de todos os pontos de partida é bastante surpreendente”. Exemplos incluem:

“Eu estava assistindo a um vídeo de soprador de folhas e três vídeos depois, era a supremacia branca”;

“Eu estava assistindo a uma discussão acadêmica sobre as origens da escravidão agrícola e o próximo vídeo foi de negadores do holocausto”;

“Eu estava assistindo a um vídeo com minhas filhas sobre Nelson Mandela e o próximo vídeo foi algo dizendo que os negros na África do Sul são os verdadeiros racistas e criminosos”.

É assustador.

Renée DiResta, especialista em desinformação e como a propaganda se espalha, notou há alguns anos que ao se juntar a um grupo anti-vacina no Facebook, o site também recomendaria grupos sobre curas naturais de câncer, rastros deixados pelas fumaças dos aviões da esquadrilha da fumaça, Terra plana e de todos tipos de grupos anti-ciência. Essas redes estão fazendo mui-

to para promover esse tipo de propaganda.

Thomas mencionou um artigo de pesquisa sobre como os laços de devolutivas descontroladas podem funcionar no policiamento preditivo. Se um software ou uma análise prever que haverá um alto índice de crimes em uma área, a polícia pode mandar mais policiais para lá - mas porque há mais policiais lá, eles podem fazer mais prisões, o que pode nos levar a pensar que há mais crimes lá, o que nos leva a enviar ainda mais policiais para lá. Podemos entrar facilmente neste ciclo de devolutivas descontroladas.

Thomas sugeriu que precisamos realmente pensar sobre a ética de incluir certas variáveis em nossos modelos. Embora possamos ter acesso aos dados, e mesmo que esses dados melhorem o desempenho do nosso modelo, é ético usar? Está de acordo com nossos valores como sociedade? Até mesmo os engenheiros precisam fazer perguntas éticas sobre o trabalho que fazem, e devem ser capazes de responder questões éticas sobre o assunto. Vamos ver menos e menos tolerância da sociedade para isso.

Angela Bassa, diretora de ciência de dados da iRobot, disse: “Não é que os dados possam ser tendenciosos. Os dados são tendenciosos. Se quiser usar dados, é necessário entender como eles foram gerados”.

Tratando o preconceito em palavras incorporadas

Mesmo se removermos o preconceito no início do desenvolvimento do modelo, existem tantos lugares em que o preconceito pode se infiltrar, que é necessário continuar procurando.

Conjuntos de dados mais representativos podem ser uma solução. Buolamwini e Gebru identificaram as falhas de preconceito nos produtos de visão computacional mencionados anteriormente e reuniram um conjunto de dados muito mais representativo de homens e mulheres com todos os diferentes tons de pele. Este conjunto de dados está disponível em [Gender Shades](#). O site também oferece o trabalho acadêmico deles e um pequeno vídeo sobre seus trabalhos.

Gebru e outros publicaram recentemente um artigo chamado [“Datasheets for Datasets”](#). O artigo fornece um conjunto de dados para registrar características e metadados que revelam como um conjunto de dados foi criado, como ele foi composto, que tipo de pré-processamento foi feito, que tipo de trabalho é necessário para mantê-lo e quaisquer considerações legais ou éticas. É muito importante entender os conjuntos de dados usados na criação dos modelos.

Thomas enfatizou que é nosso trabalho pensar em consequências não intencionais com antecedência. Pense em como certas criaturas ou assediadores ou governos autoritários poderiam usar uma plataforma que construímos. Como nossa plataforma poderia ser usada para propaganda ou desinformação? Quando o Facebook anunciou que começaria a usar sua modelagem de ameaças, muitas pessoas perguntaram por que isso não acontecia nos últimos 14 anos.

Há também um argumento para não armazenar dados de que não precisamos para que ninguém possa pegar esses dados.

Nosso trabalho é pensar em como o software pode ser mal

utilizado antes que aconteça. A cultura do campo da segurança da informação é baseada nisso. Precisamos começar a pensar mais em como as coisas podem dar errado.

Perguntas a serem feitas sobre a IA

Thomas listou algumas perguntas para perguntar sobre a IA:

- Qual preconceito está nos dados? Existe algum preconceito em todos os dados e precisamos entender o que é e como os dados foram criados;
- O código e os dados podem ser auditados? Eles são de código aberto? Há um risco quando algoritmos proprietários de código fechado são usados para decidir coisas de saúde e justiça criminal e quem é contratado ou demitido;
- Quais são as taxas de erro para os diferentes subgrupos? Se não tivermos um conjunto de dados representativos, talvez não percebamos que nosso algoritmo está tendo um desempenho ruim em algum subgrupo. O tamanho das amostras são grandes o suficiente para todos os subgrupos em seu conjunto de dados? É importante verificar isso, assim como o Pro Publica fez com o algoritmo de reincidência que analisou a corrida;

- Qual é a precisão de uma alternativa simples baseada em regras? É muito importante ter uma boa linha de base, e essa deve ser a primeira etapa sempre que estivermos trabalhando em um problema, porque se alguém perguntar se 95% de precisão é boa, precisamos ter uma resposta. A resposta correta depende do contexto. Isso surgiu com o algoritmo de reincidência, que não era mais eficaz do que um classificador linear de duas variáveis. É bom saber o que é essa alternativa simples;

- Quais processos estão em vigor para lidar com recursos ou erros? Precisamos de um processo de apelo humano para coisas que afetam a vida das pessoas. Como profissionais, temos relativamente mais poder em fazer essas perguntas em nossas empresas;

- Quão diversificada é a equipe que a construiu? As equipes que constroem a tecnologia devem ser representadas por pessoas que serão afetadas por ela, o que cada vez mais é de todos.

Pesquisas mostram que equipes diferentes têm um desempenho melhor e acreditam que somos meritocráticos, pode realmente aumentar o preconceito. Leva um tempo e esforço para fazer entrevistas de forma consistente. Uma boa referência para isso é o post do blog intitulado "[Fazendo](#)

[pequenas mudanças culturais](#)" por Julia Evans.

A tecnologia avançada não é um substituto para uma boa política. Thomas falou sobre os estudantes da fast.ai de todo o mundo que estão aplicando o aprendizado profundo a problemas sociais, como salvar florestas tropicais ou melhorar o atendimento de pacientes com mal de Parkinson.

Existem regulamentos de IA, como o Ato de Discriminação e Emprego, de 1967, e o Ato de Igualdade de Oportunidade de Crédito, que são relevantes. Estes não são perfeitos, mas são melhores do que não ter qualquer proteção, uma vez que realmente precisamos pensar sobre quais direitos, como sociedade, queremos proteger.

Thomas concluiu sua palestra dizendo que nunca pode ser caracterizado pelo preconceito. Podemos seguir alguns passos em direção às soluções, mas o preconceito pode se infiltrar em muitos lugares. Não há uma lista de verificação que assegure que o preconceito esteja em jogo e não tenhamos mais com o que nos preocupar. É algo que sempre temos que continuar procurando.



PONTOS PRINCIPAIS

- A sociedade deve exigir transparência e responsabilidade legal e financeira para o uso de algoritmos na tomada de decisão automatizada. Caso contrário, nem o público e nem uma agência reguladora serão capazes de entender ou regular algoritmos complexos e as interconexões complexas entre as redes de dados que esses algoritmos utilizam;
- Não há consenso sobre como definir, evitar ou mesmo tornar explícito o viés — distorção do julgamento — nos algoritmos usados na execução de políticas públicas ou em pesquisas científicas;
- A natureza perfeita e conveniente de muitas tecnologias, como residências personalizadas, dificulta a compreensão de onde os dados vêm, como são usados por algoritmos e para onde vão;
- Empresas e indivíduos, especialmente quando trabalham no setor público, devem assumir que os resultados das decisões dos algoritmos terão que ser explicados às pessoas que são adversamente afetadas por elas em tempo hábil, para que possam apelar ou contestar essas decisões

PODEMOS CONFIAR EM ALGORITMOS PARA TOMADA DE DECISÃO AUTOMÁTICA?

por **Michael Stiefel**

Os algoritmos subjacentes a esses sistemas podem produzir resultados incompreensíveis ou socialmente indesejáveis. Como os reguladores podem determinar a segurança ou a eficácia dos algoritmos incorporados em dispositivos ou máquinas, se não puderem compreendê-los? Como os cientistas podem entender um relacionamento baseado em uma descoberta realizada por meio de um algoritmo?

INTEGRANTES DESSE PAINEL



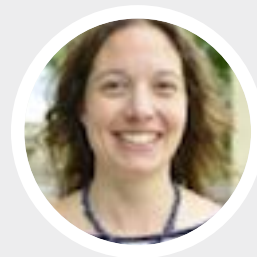
Michael Veale

é doutor e pesquisador em aprendizado de máquina e o responsável do setor público na University College London, especializada na justiça e responsabilidade de ferramentas baseadas em dados no setor público, bem como na interação entre tecnologias avançadas e lei de proteção de dados. Sua pesquisa foi citada por órgãos e reguladores internacionais, na mídia, bem como debatida no Parlamento. Ele atuou como consultor em aprendizado de máquina e sociedade para o Banco Mundial, Royal Society e British Academy, e trabalhou anteriormente em IoT, saúde e envelhecimento na Comissão Europeia. Veale pode ser encontrado no twitter em [@mikaarv](#).



Andrew Burt

é chief privacy officer e engenheiro jurídico da Immuta, uma das principais plataformas de ciência de dados e de gerenciamento de dados do mundo. Ele também é membro visitante do Projeto Sociedade da Informação da Yale Law School. Anteriormente, Burt foi consultor especial de política para o chefe da Divisão de Cyber do FBI, onde atuou como principal autor do relatório do FBI sobre o ataque de 2014 à Sony. Burt publicou artigos sobre tecnologia, história e direito no New York Times, no Financial Times, no Los Angeles Times, Slate e no Yale Journal of International Affairs, entre outros. Seu livro *American Hysteria: The Untold Story of Mass Political Extremism*, nos Estados Unidos, foi chamado de “um livro de leitura obrigatória sobre um assunto que poucos querem abordar”, do Prêmio Nobel emérito Desmond Tutu, Arcebispo Nobel. Burt é doutor em direito pela Yale Law School e é bacharel pela McGill University. Ele é membro do Conselho de Relações Exteriores, membro do Conselho de Washington, DC e da Virginia State Bars, além de coordenador de resposta a incidentes cibernéticos GIAC (Global Information Assurance Certified).



Rebecca Williams

é professora de direito público e direito penal na Universidade de Oxford. Seu trabalho inclui o exame de métodos ótimos de tomada de decisão e o uso do direito penal como forma de regulamentação. Cada vez mais seu trabalho também se concentra na relação entre lei e tecnologia e as maneiras pelas quais a lei precisará se desenvolver para acompanhar o desenvolvimento tecnológico.

Exemplos de tais áreas são: determinar quem é libertado sob fiança ou quem deverá receber crédito financeiro, prever onde ocorrerá um crime, averiguar violações das leis anti-discriminação ou julgar a culpa em um acidente com um carro autônomo.

Não está claro se os algoritmos podem detectar suas próprias falhas mais do que um ser humano pode determinar se são realmente doentes mentais. Não há

nenhuma linha de código nesses algoritmos que os instrua a fazer algo ruim a alguém.

O que podemos fazer para resolver este problema?

Integrantes deste painel:

1. **Rebecca Williams** - professora de direito público e direito penal, em associação com o Pembroke College na Universidade de Oxford

2. **Andrew Burt** - chief privacy officer e engenheiro jurídico na Immuta

3. **Michael Veale** - integrante da University College London. Departamento de Ciência, Tecnologia, Engenharia e Política Pública

InfoQ: As pessoas geralmente desconhecem o papel dos algoritmos na sociedade. Qual é a melhor maneira de educar as pessoas sobre os benefícios

A maioria das evidências úteis é causal na natureza. Queremos saber o que causa o quê e como o mundo funciona. Algoritmos de aprendizado de máquina não são tão bons nisso, e seus resultados e poder preditivo podem ser bastante frágeis como resultado.

e problemas associados ao crescente uso difundido de algoritmos?

Andrew Burt: O que mais precisamos é de história e contexto sobre como este tipo de tecnologia foi usado antes, e sobre o que é diferente agora, especialmente quando se trata do que é comumente chamado de "IA". Temos, por um lado, pessoas como Elon Musk, declarando que a IA é uma ameaça existencial à vida na Terra, que está tendo um impacto real na maneira como o público pensa sobre a IA. E temos, por outro lado, alguns defensores obstinados da IA, sugerindo que isso resolverá todos os problemas que temos. A verdade está, naturalmente, em nenhum extremo. Além disso, nem todo desafio que a AI coloca é novo. Já desenvolvemos ferramentas e práticas para enfrentar alguns desses desafios em outras áreas. Então, acho que todos se beneficiam de uma discussão mais ampla que coloque os desafios da IA em perspectiva e nos permita construir os sucessos do passado e corrigir os erros de como adotamos as tecnologias anteriores. Há muitas coisas boas que podemos fazer se acertarmos. Por outro lado, há muitos prejuízos que podem ocorrer se nos enganarmos - prejuízos discriminatórios, oportunidades perdidas e muito mais. As apostas são altas.

Rebecca Willians: Os Artigos 13(2)(f), 14(2)(g) e 15(1)(h) do [GDPR](#) declaram que os titulares de dados têm "o direito de saber a existência de tomadas de decisão automatizadas, incluindo perfis". Então, qualquer que seja a razão pela qual eles tenham acesso a informações sobre o processo, no mínimo as pessoas terão que ser avisadas quando uma decisão específica sobre elas ou a respeito delas estiver sendo tomada usando um processo automatizado. A esperança é que isso eleve a con-

sciência de quando e como esses sistemas estão sendo usados.

Em termos de educação, obviamente, quanto mais cedo começarmos com essas questões, melhor. As escolas ensinam cada vez mais a codificação aos alunos, bem como questões éticas como cidadania ou educação pessoal e social, portanto, quanto mais puder ser feito para aumentar a conscientização e discussão nesses contextos, as futuras gerações estarão mais bem preparadas quando projetarem, operarem e interagirem com esses sistemas. Isto é definitivamente algo que as universidades também podem ajudar a facilitar. Já existem contextos em que acadêmicos visitam escolas para apoiar o aprendizado e seria ótimo se isso pudesse acontecer também nesse assunto.

Isso deixa claro a questão de como podemos alcançar aqueles que passaram pela educação escolar antes que surgissem esses tipos de preocupações. Os mesmos desafios surgem aqui como surgem em relação à disseminação de qualquer tipo de informação: as pessoas tendem a confiar em certas fontes ao invés de outras, dando origem ao risco de câmaras de eco e desinformação. Haverá certamente um papel para a grande mídia aqui e balanceamento, cientificamente baseado em relatórios por esses meios será vital, como sempre, mas quanto menos confiança o público colocar em tais fontes de informação, menos eficaz será. Haverá certamente um papel para instituições como o [Gabinete do Comissário de Informação](#) para fornecer aconselhamento e informação para os cidadãos por meio do seu website, e novamente como um acadêmico gostaria de ver as Universidades ajudando também neste contexto, apoiando estes outros pontos de venda ou por meio de engajamento público direto.

Michael Veale: No design de tecnologia, vem ocorrendo uma grande tendência para tornar os sistemas “perfeitos”. Em suma, isso significa que as pessoas podem se concentrar no que desejam fazer, e não em como querem fazê-lo, o que geralmente é ótimo para os indivíduos ajudá-los a alcançar o que desejam. As casas inteligentes são um exemplo disso, embora muitas sejam um pouco desajeitadas demais para terem conquistado esse título. No entanto, com uma variedade de algoritmos de sistemas de hoje, muita uniformidade significa que os indivíduos não têm a chance de questionar se esse sistema funciona da maneira que eles querem. Sua casa inteligente pode ser personalizada, mas se não pode ver para onde e para quem está enviando os dados. Seu feed de notícias do Facebook pode parecer atraente, mas se sabe quem está sendo excluído e por quê.

Poderíamos realizar cursos sobre algoritmos na sociedade, mas é improvável que isso resolva problemas mais profundos. As tecnologias se movem rapidamente. Minha prima me contou outro dia, que na escola eles estavam aprendendo sobre segurança cibernética. “Eles nos disseram para não clicar em pop-ups”, disse ela. “Mas como vou saber como é um pop-up?”. Os navegadores mudaram muito rapidamente para bloqueá-los e, em dispositivos móveis, simplesmente não é mais o paradigma. Assim, uma educação única, a menos que esteja desenvolvendo habilidades críticas gerais, geralmente é um pouco demais para um alvo em movimento.

Assim, conseqüentemente, precisamos mesclar a educação nos produtos e serviços que usamos todos os dias. Esses serviços devem se explicar, não necessariamente com uma passagem de texto ou manual, mas em virtude de um design inteligente que deixa claro quando

fluxo de dados, decisões automatizadas e outros comportamentos estão acontecendo. Nesse caso, os indivíduos devem ser capazes de aprofundar ainda mais para ver e aprender mais, se estiverem interessados: e então, sem dúvida, sentirão melhor o que está acontecendo ao seu redor, mesmo quando as opções para perceber e detalhar não estão lá.

InfoQ: Algoritmos serão frequentemente usados na execução de políticas públicas ou em pesquisas científicas que afetarão as políticas públicas. Requisitos legais, julgamentos de valor e parcialidade são quase inevitáveis. Como os valores sociais podem ser explicitamente visíveis, e o preconceito pode ser evitado na programação do algoritmo e na interpretação dos resultados?

Burt: Do lado da tecnologia, existem todos os tipos de ferramentas importantes que estão sendo desenvolvidas para ajudar a minimizar muitas dessas desvantagens. Uma ferramenta chamada [LIME](#), que ajuda a explicar os chamados algoritmos de caixa preta, é um ótimo exemplo. Um cientista de dados chamado Patrick Hall realmente merece um elogio por fazer um [ótimo trabalho](#) sobre interpretabilidade no aprendizado de máquina. E há muitos outros exemplos para citar. Nossas equipes de engenharia jurídica e ciência de dados estão se mantendo no topo de todos esses desenvolvimentos na Immuta.

Mas acho que muitas vezes é esquecido o lado processual. Os processos usados para desenvolver e implantar Machine Learning (ML) são incrivelmente importantes, e modelar estruturas de gerenciamento de risco como o [SR 11-7](#) do Federal Reserve Board que há muito tempo já reconhecem esse fato. Esse regulamento se

aplica ao uso de algoritmos em instituições financeiras nos EUA. O pessoal do AI Now Institute também avançou com o que eles chamam de [avaliações do impacto de algoritmos](#), que oferecem outra estrutura para esse tipo de abordagem.

Há muita coisa lá, francamente, e lançaremos um white paper resumindo brevemente algumas dessas práticas recomendadas - técnicas e processos - para ajudar nossos clientes e os outros a gerenciar os riscos da implantação de modelos de aprendizado de máquina na prática. Estamos trabalhando duro para finalizar o white paper e estamos animados para lançá-lo nos próximos meses.

Willians: Existem várias maneiras diferentes de abordar essa questão. Primeiro, é vital examinar cuidadosamente os dados usados para treinar e operar sistemas automatizados de tomada de decisão. Se os dados em si forem tendenciosos, o resultado também será. Vem ocorrendo muita discussão sobre os sistemas de previsão de risco usados no contexto da justiça criminal em vários estados dos EUA e a dificuldade com esses sistemas é que eles tendem a superestimar a reincidência de réus negros ao mesmo tempo que a subestimam para réus brancos. Mas, apenas para dar um exemplo, um potencial preditor de risco usado pode ser a prisão antecipada por ofensas de posse menores. E ainda assim tais ofensas são mais prováveis de serem detectadas por stop e search, e as táticas de stop e search tendem a se inclinar na mesma direção: prever uma razão para parar e procurar pessoas negras enquanto prevê a necessidade de parar e procurar pessoas brancas. Então, como parar e pesquisar é distorcido contra os negros em favor dos brancos, mais pessoas negras são consideradas portadoras do que brancas e, assim, os negros são calculados a ter

um risco maior de reincidência do que os brancos. A discriminação inicial na coleta de dados, alimenta todo o sistema na saída. Portanto, se acharmos que nossos dados iniciais provavelmente produzirão esse tipo de efeito distorcido, devemos pensar cuidadosamente sobre se é ou não apropriado usá-lo, e talvez precisemos pensar em impor direitos para coletar dados de contrapeso.

Em segundo lugar, há importantes escolhas políticas a serem feitas no processo de codificação do sistema. O trabalho de [Krishna Gummadi](#) mostrou que nem sempre é possível ter um bolo e comê-lo. Normalmente, será necessário escolher entre diferentes medidas de precisão. Assim, por exemplo, um sistema que tem o método mais preciso de predição no agregado, considerado em todos os casos, também pode ter o maior problema de produzir resultados distorcidos em relação a categorias específicas de casos, como as mencionadas anteriormente. Ou, inversamente, um sistema que tenha precisão máxima em relação a qualquer categoria específica (como status étnico ou gênero) pode não ter um grau tão alto de precisão em todas as categorias em conjunto. É vital que tais escolhas políticas entre sistemas diferentes sejam entendidas como sendo apenas isso; são escolhas de políticas que devem ser feitas de forma aberta e transparente e por uma entidade que pode ser responsabilizada por fazê-las, não inconscientemente, por codificadores anônimos.

Terceiro, mesmo que estejamos confiantes de que fizemos tudo o que podemos, [ex ante](#) (antes de um evento ocorrer — termo jurídico) para coletar dados balanceados e fazer escolhas de códigos responsáveis, também será necessário um [ex post](#) (após o evento ocorrer — termo jurídico) para assegurar, que tais siste-

mas, sejam sujeitos a auditorias regulares para assegurar que eles não estejam espontaneamente gerando formas de discriminação que não havíamos previsto. Será necessário fazer isso mesmo se não tivermos certeza do motivo pelo qual isso está acontecendo, mas, em quarto lugar, também é vital que façamos tudo o que pudermos para tornar os algoritmos transparentes e responsáveis, de modo que, se uma auditoria deste tipo detectar um problema, possamos ver onde e como aconteceu. Há um número de pessoas trabalhando nisso e um grupo nosso em Aberdeen (Prof. Pete Edwards), Oxford e Cambridge (Dr. Jat Singh) acabam de [receber uma doação do EPSRC](#) para trabalhar mais nessa questão.

Em termos das fontes de regulação para cada uma dessas quatro questões, os sistemas serão usados por entidades públicas e privadas. Onde eles são operados por entidades públicas ou governamentais, acho que há definitivamente um papel para o direito público existente de desempenhar na responsabilização de tais entidades e impor mais deveres de transparência, justiça, etc., que já são inerentes ao direito público. Para as entidades privadas, o desafio será pensar quais desses deveres de transparência, responsabilidade e justiça devem ser levados para o setor privado, como o preço pelo aumento de poder oferecido por tais sistemas.

Veale: A maioria das evidências úteis é causal na natureza. Queremos saber o que causa o quê e como o mundo funciona. Algoritmos de aprendizado de máquina não são tão bons nisso, e seus resultados e poder preditivo podem ser bastante frágeis como resultado. A principal maneira de tornar os valores sociais explicitamente visíveis é desacelerar e reconhecer que nossos objetivos muitas vezes não são apenas pre-

visão, mas compreensão. Estamos em grande perigo de treinar uma geração de pessoas que podem fazer o primeiro, mas não o segundo. Quando construímos modelos causais, temos uma oportunidade maior de discutir se é assim que queremos que o mundo funcione e se comporte. Talvez seja, talvez não seja: mas é uma conversa que é mais visível e muito mais fácil de ter e de comunicar.

InfoQ: Em maio deste ano, o Regulamento Geral de Proteção de Dados da União Europeia (GDPR) entra em vigor. Entre suas disposições, está o Artigo 22, que trata da tomada de decisão individual automatizada. Muitas pessoas argumentam que essa regra exige não apenas que os direitos de privacidade dos dados sejam respeitados, mas que as decisões tomadas pelos algoritmos sejam explicáveis.

O que acha dessa interpretação do regulamento? Este regulamento exige que os dados sejam removidos do uso por algoritmos? Se sim, isso poderia reduzir a eficácia do algoritmo? Em geral, a abordagem da União Europeia é válida, ou a “lei das consequências não intencionais” vai piorar a situação?

Burt: Há um enorme debate em andamento na comunidade jurídica sobre como, exatamente, o GDPR afetará a implantação do aprendizado de máquina. E dado que o GDPR só entrou em vigor em Maio de 2018, ainda há muita coisa no ar. Mas a minha opinião é que o Artigo 22 precisa ser lido ao lado dos Artigos 13-15, que afirmam que os titulares de dados têm o direito de “informações significativas sobre a lógica envolvida” em casos de tomada de decisão automatizada. Na prática, acho que isso significará que os titulares de dados terão o direito de ser in-

struídos sobre quando, por que e o mais importante, como algo como um modelo de aprendizado de máquina está usando seus dados. Como acontece com qualquer análise legal, há uma tonelada de nuances aqui. Por isso, incentivo os leitores a verificarem um [artigo anterior](#) que coloquei sobre o assunto para a Associação Internacional de Profissionais de Privacidade. Também vale a pena mencionar que um grupo chamado Working Group 29, que tem uma enorme influência sobre como as leis de privacidade da UE são aplicadas, saiu com suas próprias orientações sobre este assunto, [afirmando categoricamente](#) que a tomada de decisão automatizada é proibida por GDPR, com certas isenções.

Willians: Já sabemos que há um intenso debate entre Goodman e Flaxman, que [argumentam que o GDPR dá um “direito à explicação” completo](#), enquanto Wachter, Mittelstadt e Floridi, na minha opinião, de forma mais plausível, [argumentam que isso será suficiente para dados sujeitos a um componente de aprendizado de máquina a ser informado da existência e quais medidas de precisão estão sendo usadas para checá-lo](#). Concorro com eles que o assunto dos dados deve ser informado mais do que apenas quais pontos de dados estão sendo usados, mas também como eles são ponderados nas circunstâncias. Como mencionei anteriormente, em que o sistema está sendo operado por uma entidade pública, acho que existe um potencial significativo para uma analogia a ser tirada com nossa abordagem atual para as decisões do [Procedimento de Material Fechado](#), no qual o impacto sobre o indivíduo é significativo, ele/ela tem o direito de saber, pelo menos, a ‘essência’ do processo contra ele/ela, de modo que ele/ela possa fazer uso ‘significativo’ do direito de resposta. Isso pode

envolver apenas uma explicação ex ante, como sugerem Wachter, Mittelstadt e Floridi, mas também pode incluir explicações ex post. Em relação às entidades privadas, a situação é mais difícil, uma vez que estão geralmente sujeitas a menos deveres, embora a nossa lei existente sobre discriminação faça algum trabalho e haja também a possibilidade de deveres de estilo público serem associados ao uso de tais sistemas, mesmo em um contexto privado.

O Art 17 permite o direito de apagar dados pessoais, mas não onde o processamento é necessário para cumprir uma obrigação legal. A principal distinção aqui é entre dados individuais e gerais. Para a remoção de dados individuais, há alguns direitos limitados, como do Art 17, mas para qualquer dever ou obrigação de remover dados gerais (ou seja, dados que afetam uma categoria inteira de pessoas, como os dados descritos anteriormente), pode ser necessário olhar para as disposições mais gerais no regulamento, como “medidas adequadas para salvaguardar os direitos e liberdades e interesses legítimos da pessoa em causa”, ou deveres gerais em, por direito público (onde que processa dos dados é uma entidade público/governamental) ou lei que proíbe a discriminação.

Novamente, isso depende, se o que está sendo usado é dado individual ou geral. A remoção de dados gerais distorcidos pode tornar o algoritmo mais preciso, ao passo que a remoção de dados individuais precisos em relação a tipos específicos de candidatos pode torná-lo mais impreciso e dar origem ao efeito de distorção.

Não acho que alguém sabe a resposta para isso com certeza neste momento! Acho que será necessário lembrar as auditorias ex post discutidas anteriormente neste painel, de modo que, se na

prática vemos consequências não intencionais, há uma oportunidade de pegá-las e resolvê-las.

Veale: O Artigo 22 no GDPR é uma disposição realmente antiga. Ela remonta à lei francesa de 1978, e boa parte dela permanece inalterada em relação ao artigo 15 da Diretiva de Proteção de Dados em 1995 (Lei de Proteção de Dados do Reino Unido de 1998). No entanto, não tem sido muito utilizado, e alguns acadêmicos o chamaram de “direito de segunda classe” como resultado.

O propósito fundamental do Artigo 22 é garantir que, se uma organização quiser tomar uma decisão totalmente automatizada e potencialmente significativa sobre alguém, ela precisa ter uma base legal para fazê-lo (consentimento livre, necessidade de executar um contrato ou obrigação legal). Se a organização não tiver um desses, eles não poderão tomar a decisão. Se garantirem um, eles têm que colocar salvaguardas em prática para garantir que a decisão seja tomada de forma justa, incluindo permitir que um indivíduo desafie a decisão. Não está claro em muitos casos como esse desafio funcionará: muitas decisões importantes são tomadas rapidamente. Se um vídeo de um evento político, tópico, for automaticamente removido do Youtube, com que rapidez ele poderá ser reativado? Se o tempo de relevância tiver passado, uma revisão humana é de pouca utilidade.

Outra destas salvaguardas, para além do desafio humano, é descrita no Recital 71 do GDPR. Os recitais, que começam uma lei europeia, destinam-se a ilustrar seu espírito e contexto, mas em leis muito disputadas como o GDPR, tornaram-se, frustrantemente para os advogados, um lugar para colocar coisas que realmente deveriam estar nos principais artigos obrigatórios. Esta salvaguarda da

explicação, ao contrário de outras, como o direito à intervenção humana, foi colocada lá, e assim veremos se e quando o Tribunal de Justiça Europeu acha que é obrigatório para os responsáveis pelo tratamento de dados.

No entanto, não vamos esquecer o significado real do Artigo 22, que não é apenas sobre explicações. Isso definitivamente restringe alguns usos de algoritmos de sistemas que as pessoas acreditam que são injustos. Contratação automatizada e filtragem de Curriculum, por exemplo, são técnicas que são altamente suspeitas nos termos do Artigo 22. Quando se está decidindo entrevistar alguém automaticamente, usando um dos produtos analíticos no mercado hoje, estamos provavelmente tomando uma decisão apenas automatizada e significativa. Qual é a sua base legal? Não há um contrato e, provavelmente, não tem uma obrigação legal, o que permite o consentimento. O ato de consentir algo automático em qualquer contexto de emprego é altamente problemático devido aos desequilíbrios de poder, e raramente pode ser visto como dado livremente. Pessoalmente, penso que o Artigo 22 torna muitas práticas automáticas de contratação em larga escala muito legalmente suspeitas.

InfoQ: Qual é a questão crítica que as sociedades enfrentam com o uso generalizado de algoritmos em vez de humanos para tomar decisões críticas?

Burt: Em duas palavras: falhas silenciosas. À medida que começamos a nos basear mais em algoritmos complexos, especialmente em várias formas de redes neurais, nossa capacidade de explicar seu funcionamento interno se tornará progressivamente mais difícil. Isso não é simplesmente porque esses modelos são difíceis de interpretar, mas porque as redes às quais

os conectamos estão se tornando cada vez mais complexas. Todos os dias, o mundo da TI fica mais difícil de gerenciar, temos mais endpoints, mais dados, mais bancos de dados e mais tecnologias de armazenamento do que nunca. E assim acredito que nosso maior desafio está em entender os ambientes de dados, nos quais estamos confiando. Porque, se não o fizermos, existe uma possibilidade muito real de estarmos constantemente a confrontar falhas silenciosas, em que algo correu mal e que simplesmente não sabemos, com consequências muito reais, e potencialmente devastadoras.

Willians: Acho que a maioria das pessoas iria encapsular isso na palavra "justiça". Mas isso realmente se resume em transparência e responsabilidade: (1) precisamos saber o máximo possível sobre o que esses sistemas estão fazendo, como e por quê. (2) É necessário haver uma entidade apropriada para responsabilizar-se por eles e um sistema apropriado e acessível para responsabilizar essa entidade.

Nossas estruturas legais e reguladoras precisam fornecer e incentivar essas duas coisas, trabalhando em estreita colaboração com os cientistas da computação que geram os sistemas.

Veale: O maior problema aqui é que os algoritmos exigem manutenção e supervisão, o que pode ser difícil de fazer em pequena escala. Eles teoricamente permitem um enorme volume e velocidade de decisões automatizadas, muito mais do que um ser humano pode fazer. Pequenas organizações podem realmente se beneficiar disso. Anteriormente, se as organizações queriam que muitas decisões acontecessem, precisavam de muita gente. Essas pessoas poderiam fornecer supervisão e feedback, mesmo que trouxessem seus próprios preconceitos. Agora,

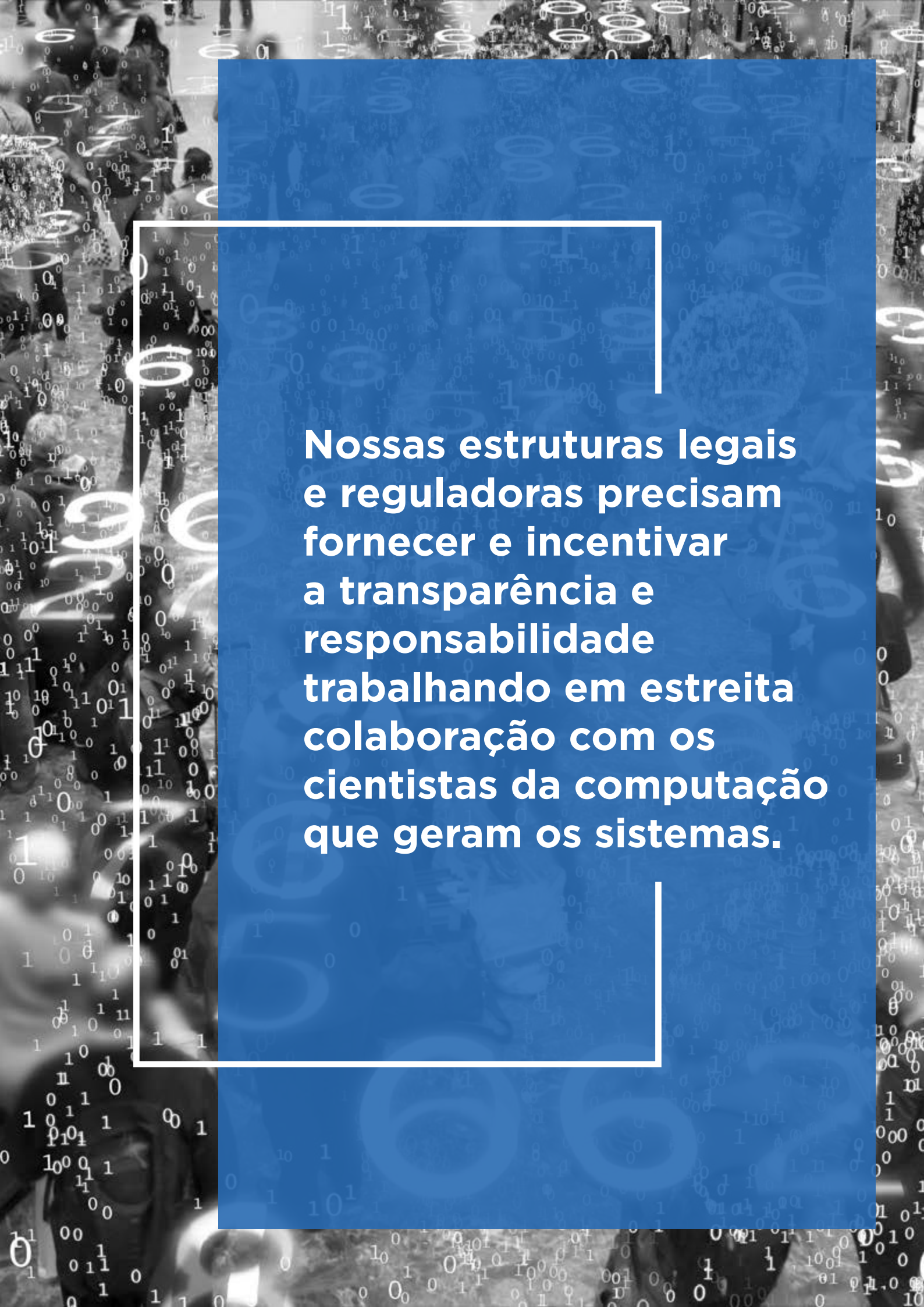
algumas pessoas podem implantar e gerenciar enormes infraestruturas de tomada de decisão, mas elas não trazem a capacidade humana de examiná-las e mantê-las. Isso cria um enorme desequilíbrio, particularmente para organizações de baixa capacidade que podem ser tentadas por confiar na automação e no aprendizado de máquina. Nestes casos, a supervisão externa é necessária; mas quem fornece isso? Quem paga por isso? E como isso realmente se encaixa em alguns desafios ocultos que a tomada de decisões algorítmicas pode causar, desafios que são frequentemente enterrados nas organizações e em suas políticas de trabalho?

Conclusão

Deixar de levar em consideração o que o público teme ou a incapacidade de prever consequências adversas impediu tecnologias como energia nuclear e culturas geneticamente modificadas.

A cidade de Nova York está estabelecendo uma [força-tarefa para propor recomendações](#) para explicações e mitigações para as pessoas afetadas pelo uso de algoritmos pelas agências da cidade. O Regulamento [Geral de Proteção de Dados da União Européia](#) é outra tentativa de começar a lidar com a questão.

Carl Jung tem a fama de ter dito que dentro de cada ser humano esconde-se um lunático. Se os algoritmos modelam o comportamento humano, o que isso significa para a sociedade?



**Nossas estruturas legais
e reguladoras precisam
fornecer e incentivar
a transparência e
responsabilidade
trabalhando em estreita
colaboração com os
cientistas da computação
que geram os sistemas.**

Edição anterior:



Lançada no QCon São Paulo 2019, a eMag Ética na Tecnologia é uma série de artigos que se propõe a entender como profissionais do desenvolvimento de software pensam sobre ética e de quem é a responsabilidade de tomar medidas razoáveis para garantir que os produtos de tecnologia não prejudiquem as pessoas.

[Clique aqui e faça o download](#)