Data Science em Direito: uma Introdução

Alexandre Costa e Henrique Costa

Data Science em Direito: uma Introdução

Alexandre Costa e Henrique Costa

Esse livro está à venda em
http://leanpub.com/data_science_direito

Essa versão foi publicada em 2022-02-15

Esse é um livro Leanpub. A Leanpub dá poderes aos autores e editores a partir do processo de Publicação Lean. Publicação Lean é a ação de publicar um ebook em desenvolvimento com ferramentas leves e muitas iterações para conseguir feedbacks dos leitores, pivotar até que você tenha o livro ideal e então conseguir tração.

This work is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International License

COSTA, Alexandre; COSTA, Henrique. Data Science em Direito: uma Introdução. Disponível em: novo.arcos.org.br/datascience_e_direito. Acesso em: 15/02/2022.

Conteúdo

Data Science em Direito: uma Introdução

1. Ciência de dados?

Ciência é uma palavra com significados múltiplos e fugidios, que nos conduzem a uma série de debates complexos.

> Será o direito uma ciência?
>
> Faz sentido usar a mesma palavra para tratar das ciências sociais e das ciências exatas?
>
> Devemos confiar verdadeiramente nos conhecimentos ditos científicos?

Falar de Ciência de Dados nos coloca no meio dessa rede de questionamentos sobre as fronteiras e as potencialidades do conhecimento científico. Inobstante, a expressão Data Science se consolidou como título que designa uma série de abordagens que lidam com o desafio de compreender a multidão de dados a que tivemos acesso ao longo do século XXI.

A disciplina que se designa por data science não pode ser considerada uma ciência, no sentido estrito desta palavra. Física e Sociologia, por exemplo, são ramos do conhecimento humano que tratam de um objeto empírico determinado e que são compostos por discursos que tentam descrever e explicar esses objetos a partir da observação cuidadosa de fenômenos empíricos.

O conhecimento científico é sempre na observação de fenômenos particulares, em busca de se compreender os padrões envolvidos

em tais ocorrências, de forma que seja possível elaborar explicações para as situações observadas. Isso vale para a física, para a biologia ou para a sociologia: observamos fenômenos, descrevemos essas ocorrências a partir de uma rede de categorias e buscamos compreendê-los por meio da formulação de explicações.

A maior parte de nosso conhecimento decorre da observação de fenômenos naturais ou sociais que não são planejados pelos pesquisadores: os movimentos dos astros, as guerras, as revoluções, as epidemias. Em certas áreas particulares, é possível construir artificialmente a situação a ser observada: físicos constroem aceleradores de partículas, médicos tratam algumas pessoas com um fármaco para avaliar sua eficiência como remédio, psicólogos colocam certas pessoas frente a situações simuladas para avaliar a sua reação. Essas situações artificialmente criadas são chamadas de experimentos e vários campos científicos, como a medicina, se desenvolvem principalmente a partir de abordagens experimentais.

A categoria Data Science deveria nos causar o mesmo estranhamento que seria gerado caso alguém formulasse a categoria Experimental Science para designar o conhecimento ligado à formulação e interpretação de experimentos. Não se trataria de designar algumas abordagens científicas como experimentais, mas a de postular a existência de uma ciência geral sobre os experimentos.

Creio que a maioria de nós rejeitaria essa proposta por considerar que a realização de experimentos é uma estratégia de pesquisa, comum a diferentes áreas, mas os saberes envolvidos no planejamento e na execução de experimentos não devem ser confundidos com o conhecimento científico gerado a partir dessas abordagens.

A formulação, a aplicação e a interpretação de experimentos são parte do que tradicionalmente chamamos de metodologia científica, uma matéria que abordamos em outro curso[1]. As abordagens observacionais envolvem esforços semelhantes aos realizados por

[1] https://novo.arcos.org.br/datascience_e_direito/metodologia.arcos.org.br

pesquisadores experimentais: coleta de dados, tratamento das informações, identificação de padrões, análises dos dados produzidos pela descrição linguística dos fenômenos observados.

Tanto as estratégias observacionais quanto as experimentais envolvem a produção e a análise de dados. Portanto,

> toda ciência sempre foi uma ciência de dados!

Cada pesquisa científica observa muitos fenômenos particulares, cataloga dados sobre eles, formula classificações e analisa os resultados alcançados. Todo pesquisador coleta informações e busca estabelecer padrões que possibilitem o desenvolvimento de modelos descritivos (explicando como fenômenos ocorrem) e/ou modelos explicativos (explicando os motivos pelos quais certos fenômenos ocorrem).

As estratégias envolvidas na chamada data science envolvem elementos que são tipicamente estudados na academia em disciplinas ligadas às metodologias quantitativas: estatística, amostras, inferências, regressões, hipóteses. Isso faz com que a data science esteja muito distante do conceito típico de ciência e bem próxima do que se descreveria normalmente como um conhecimento metodológico: um data scientist não é uma pessoa que conhece dados e suas formas de interação, mas é uma pessoa capacitada a realizar inferências complexas a partir de conjuntos de observações.

Assim, data science é melhor descrita como uma competência (uma capacidade de realizar certas atividades) do que como uma ciência (no sentido de um conhecimento rigoroso acerca de um objeto determinado). Signo desse caráter instrumental da "ciência de dados" é o fato de que se espera que todo cientista de dados tenha um sólido conhecimento material acerca dos objetos estudados.

É compreensível que todo campo de conhecimento deseje adotar a denominação de ciência, pois essa é a forma de conhecimento com maior peso nas sociedades modernas e contemporâneas. A

dogmática jurídica, há tempos, tenta (sem muito sucesso...) afirmar-se como uma ciência, apesar das contínuas críticas dirigidas a essa pretensa cientificidade do direito.

Ao reivindicar o título de cientistas, as pessoas capacitadas a lidar com grandes quantidades de dados parecem ingressar numa busca de prestígio semelhante ao dos pretensos "cientistas do direito". Porém, apesar de se tratar de uma designação problemática, o rótulo "data science" tem tido amplo reconhecimento e utilização, o que indica que esse tipo de habilidade tem sido especialmente valorizada no mundo contemporâneo.

Para além dos elementos de marketing e de valorização social de seus conhecimentos, devemos reconhecer que o avanço da data science é consequência de um fenômeno radicalmente contemporâneo: a existência de uma quantidade imensa de dados disponíveis, que não foram processados.

Até recentemente, nós éramos capazes de processar todos os dados que produzíamos, ou seja, as informações que éramos capazes de coletar. Os cientistas desenvolvem modelos explicativos, mas para isso eles precisam de dados, ou seja, de registros que ofereçam informações sobre o mundo. Os fenômenos ocorrem no mundo e não conseguimos observá-los todos ao mesmo tempo.

Nosso cérebro opera por meio de vieses e de heurísticas que lhe permitem atuar de forma competente, mesmo quando dispõe de poucos dados. De fato, nosso sistema nervoso opera melhor em face a poucos dados, já que a multiplicação de informações desafia nossas capacidades cognitivas: nossa memória é pequena e nossa capacidade de processamento cerebral é muito limitada para realizar cálculos matemáticos e estatísticos. Nossa espécie sobreviveu milênios com base nessa nossa capacidade observar padrões de forma intuitiva, gerando conhecimentos que decorrem da generalização de poucas experiências.

Quando desenvolvemos a habilidade de registrar informações, especialmente de registrar medidas, nós nos encontramos frente à

possibilidade de lidar simultaneamente com uma multiplicidade de registros que seria impossível processarmos de outra forma. A escrita cuneiforme dos antigos sumérios, que é o sistema mais antigo de escrita que conhecemos, servia justamente a essa função: armazenar quantidades numéricas. Como indica Yuval Harari:

> A escrita é um método para armazenar informações por meio de símbolos materiais. O sistema de escrita sumério fez isso combinando dois tipos de símbolos, que eram gravados em pequenas tábuas de argila. Um tipo de símbolo representava os números. Havia símbolos para 1, 10, 60, 600, 3600 e 36000 (os sumérios usavam uma combinação de sistemas numéricos de base 6 e de base 10. Seu sistema de base 6 nos deixou vários legados importantes, como a divisão do dia em 24 horas e do círculo em 360 graus). O outro tipo de símbolo representava pessoas, animais, mercadorias, territórios, datas e assim por diante. Ao combinar ambos os tipos de símbolos, os sumérios foram capazes de preservar muito mais dados do que qualquer cérebro humano poderia se lembrar ou qualquer cadeia de DNA poderia codificar.

O cérebro continuava sendo uma máquina pobre para lidar com números, mas a existência de uma multidão de registros possibilitava que nós contássemos com uma espécie de memória auxiliar, oferecida pela escrita. A partir de então, nosso desafio maior não era o de lembrar, mas o de organizar os registros, tornando as informações acessíveis e buscando encontrar nelas padrões que não conseguiríamos captar na observação direta dos fatos.

A ciência moderna, tal como a astronomia dos antigos, é baseada nessa observação cuidadosa de fenômenos complexos, medindo suas diversas dimensões (tempo, espaço, intensidade, etc.) e, com isso, tornando possível uma descrição quantitativa do mundo que, por sua vez, possibilita a utilização de cálculos matemáticos como formas de descrição dos acontecimentos.

Para que Kepler pudesse formular sua teoria acerca da órbita elíptica dos planetas, ele precisou contar com muitos (mas muitos mesmo!) registros de observações particulares, feitas de forma precisa e cuidadosa, ao longo de anos de trabalho. A meticulosidade de Tycho Brahe, em observar fenômenos e catalogar dados, foi um requisito fundamental para o desenvolvimento das teorias de Kepler. Hoje, os aceleradores de partículas e os telescópios fornecem aos físicos uma multidão de dados que, uma vez processados, podem revelar padrões que não enxergávamos.

As observações naturais e os experimentos são máquinas de gerar dados, e o elemento mais complexo e criativo das ciências está justamente na compreensão desses dados: na organização das informações, na classificação dos fenômenos, na formulação de hipóteses explicativas. Já não dependemos mais do barro da escrita cuneiforme, dos papiros egípcios nem do papel em que Kepler fez suas anotações. Todavia, as ciências continuam enfrentando esse desafio ancestral: extrair conclusões sólidas a partir de uma multiplicidade de informações meticulosamente registradas.

2. Pesquisa x Dogmática

A abordagem indutiva das ciências, que formulam explicações gerais a partir de uma multiplicidade de observações particulares, não corresponde à atividade típica dos juristas.

Juristas não são pesquisadores, que produzem modelos descritivos e explicativos a partir da observação cuidadosa de fenômenos. Operadores do direito são técnicos que produzem discursos retóricos, voltados a interferir no comportamento de outras pessoas ou a tomar decisões. O discurso jurídico é dogmático e não científico, sendo que a principal diferença entre esses dois tipos de abordagem está na forma como o conhecimento é gerado.

A ciência decorre de pesquisa e a pesquisa é uma observação direta

da realidade. Os juristas, por sua vez, operam um saber dogmático que não é resultado de pesquisas empíricas, mas de um outro processo: a acumulação cultural. Os cânones de hermenêutica, os princípios jurídicos e os conceitos da teoria jurídica não decorrem de uma observação de fatos empíricos, mas de certas estruturas semânticas que se consolidam em uma determinada cultura.

Os juristas não precisam saber como os fenômenos efetivamente ocorrem porque a sua atividade é baseada na crença de que a atuação da burocracia judicial seguirá determinadas estruturas argumentativas que são percebidas como adequadas pela comunidade dos juristas. O interessante é que essa crença parece relativamente justificada pelos fatos: aparentemente, a prática judicial segue, ao menos em grande medida, os cânones definidos pela dogmática.

Por que isso acontece? Provavelmente porque a educação dos juristas proporciona o compartilhamento de certas crenças e de certos modelos descritivos, que fazem com que a sua atuação seja efetivamente norteada pelos discursos dogmáticos. O curioso é que nós sabemos que cada jurista, em particular, é uma pessoa com preferências ideológicas particulares, com interesses sociais diversos, com heurísticas próprias e com vieses de confirmação que condicionam suas conclusões particulares. Não obstante, essa imensa variação individual é equilibrada pelo fato de que a educação dos juristas gera uma espécie de discurso comum, com regras de aplicação relativamente estáveis.

Por mais que a atuação individual de uma pessoa seja imprevisível, a solução das questões jurídicas envolve uma atuação coordenada de várias pessoas (advogados, juízes, ministério público, desembargadores, ministros), que normalmente resultam em soluções relativamente compatíveis com os cânones da dogmática. Não importa que essa dogmática seja uma construção cultural arbitrária, nem que ela esteja assentada sobre descrições ideológicas e categorias mistificadoras. O mais importante é que ela serve como campo no qual se articulam as várias subjetividades.

A dogmática, não por acaso, opera como o discurso religioso: de nada importa saber se determinados deveres são objetivamente sagrados, quando as pessoas de uma comunidade vivem esses deveres como se eles fossem sagrados. O mais curioso é que a observação das condutas não nos permite diferenciar a crença sincera e acrítica nos cânones dogmáticos de uma abordagem que se apropria retoricamente desses cânones. O resultado final é idêntico: produzimos petições e decisões que argumentam como se a dogmática fosse válida e criamos uma rede de expectativas sociais que é estabilizada pela reprodução desses cânones, que tornam as interações mais previsíveis e a vida social mais estável.

Existe, inclusive, uma tendência de perceber como técnicas as decisões que aparentemente seguem as orientações da dogmática dominante e como políticas as decisões que rompem esses critérios. Por mais que a filosofia da linguagem do século XX nos tenha tornado (ou ao menos nos devesse ter tornado) atentos para o fato de que não existem decisões jurídicas puramente técnicas, essa distinção continua sendo utilizada na compreensão atual do comportamento das cortes.

Cem anos depois da teoria pura do direito, continua causando estranhamento ao senso comum a tese de que não existem critérios técnicos capazes de justificar a escolha entre duas opções interpretativas plausíveis e de que toda aplicação judicial do direito deve ser entendida como o exercício de uma opção político-ideológica.

A admissão do caráter político da atuação judicial, em tese, deveria ter conduzido os juristas a realizar pesquisas empíricas, voltadas a compreender como os juízes (e demais juristas) atuam, quais são as suas respostas a certos argumentos, quais são as formas pelas quais eles efetivamente formulam as suas teses e sentenças. Mas não foi isso o que ocorreu ao longo do século XX, pois continuamos ligados à reprodução de um discurso dogmático, educando os juristas para conseguirem identificar as soluções corretas, e não os padrões efetivos de comportamento judicial.

Por operar no nível dos cânones dogmáticos, os juristas são dispensados de fazer uma análise psicológica dos motivos dos juízes, uma análise sociológica dos interesses cristalizados nas categorias civilistas, uma análise política de quem é beneficiado pelas teorias dominantes. Essa é uma situação que permite que a operação de certas estruturas retóricas garantam uma grande efetividade dos discursos jurídicos, sem a necessidade de desenvolver um conhecimento aprofundado sobre os fatos.

O reconhecimento de que a atividade jurídica é política não alterou substancialmente a prática dessa atividade. Como Kelsen diagnosticou, os juristas vivem dentro de uma ficção: a ficção de que as normas são válidas e que o discurso dogmático efetivamente guia a prática dos juristas. Como os juristas vivem essa ficção como realidade, terminamos como os habitantes da caverna da alegoria de Platão: quem confunde a ficção com a realidade tem uma capacidade peculiar de interferir no comportamento das pessoas que compartilham a mesma ficção.

O discurso dogmático não é operativo e eficiente apesar da sua ficcionalidade. Ele é operativo e eficiente por causa dessa ficcionalidade: ele opera independentemente de sua relação com a empiria. Portanto, o fato de os juristas não serem pesquisadores não decorre de uma especial ignorância, nem de uma falha de sua formação. Decorre apenas do reconhecimento (ao menos intuitivo) de que é possível alcançar patamares de eficiência adequados, sem a necessidade de elaborar um conhecimento científico que é extremamente caro e trabalhoso.

Imagine que um advogado seja capaz de conseguir resultados mais eficazes se ele tiver um conhecimento preciso e meticuloso sobre a atuação de cada juiz, sobre a atuação de cada assessor, sobre as preferências estilísticas de cada pessoa envolvida em um julgamento, sobre os métodos de trabalho de cada gabinete, sobre todos os elementos que poderiam influenciar em uma decisão. Será que o custo de levantar todas essas informações para cada processo particular valeriam os eventuais benefícios? E será que

esses elementos concretos seriam efetivamente úteis, dado o alto grau de imprevisibilidade de um provimento judicial que depende de múltiplos eventos aleatórios?

Pode ser que, de fato, a aposta na produção de cânones dogmáticos estáveis proporcione uma prática jurídica mais previsível do que o conhecimento e a regulação dos fatores empíricos subjacentes: as crenças individuais dos julgadores, os incentivos dos assessores, as ideologias sociais dominantes, etc. O fato de que a dogmática jurídica (um discurso tecnologicamente voltado a organizar uma prática decisória) continua sendo a base da formação dos juristas sugere que essa forma de abordagem ainda é a mais capaz de gerar resultados eficazes.

Essa formação dogmática (muitas vezes motivada por uma crença acrítica na veracidade da dogmática) faz com que os juristas tendam a orientar suas ações por esses repertórios normativos, que dizem quais são os conceitos adequados e as formas corretas de interpretar a lei e decidir os casos. Isso faz com que os juristas do século XXI sejam parecidos com os médicos do séx. XVIII: dominam bem o repertório de conhecimentos compartilhados e sabem prever os diagnósticos que os outros médicos farão, o que lhes permite fazer diagnósticos que serão bem recebidos pela comunidade dos médicos. Para alcançar prestígio no campo, não importa muito se os diagnósticos são corretos ou equivocados, mas apenas se eles são aceitos pelos demais operadores da medicina ou do direito.

No caso da medicina, o saber tradicional de médicos experientes foi sendo aos poucos substituído pelo saber científico, decorrente de pesquisas controladas, de observações que seguem métodos claros. No caso do direito, o conhecimento científico não teve grande impacto porque o resultado de um processo judicial é (ou deveria ser) definido por decisões tomadas por pessoas que operam o mesmo discurso dogmático utilizado pelos advogados que redigem as petições e também porque esses resultados seguem padrões menos previsíveis que as reações biológicas mapeadas pela medicina.

No caso dos médicos, conhecer bem as teorias vigentes e as concepções dominantes entre os pares não é uma garantia de que os tratamentos terão sucesso, pois a doença segue seu curso independentemente do consenso dos médicos. Tragicamente, os consensos médicos acerca dos modos de transmissão de doenças viróticas pulmonares fizeram com que tardasse muito o reconhecimento de que, contrariando os consensos estabelecidos, o SarsCov2 tem os aerossóis como forma predominante de contágio. Porém, o avanço das pesquisas médicas tem alterado as estratégias de enfrentamento da pandemia em que estamos imersos enquanto escrevemos este texto.

Saber aplicar as teorias médicas dominantes é importante porque esse tipo de conhecimento permite a cada médico apropriar-se de um repertório de conhecimentos tradicionais, cuja manutenção no senso comum está normalmente está ligada a uma razoável eficácia de seus diagnósticos e terapêuticas.

No caso dos juristas, a aplicação dos saberes tradicionais continua sendo a estratégia mais eficiente, pelo fato peculiar de que o objetivo não é conhecer os fatos, mas influenciar na decisão de processos que deveriam ser decididos com base na dogmática hegemônica. A capacidade de atuar com eficiência na prática jurídica não é desenvolvida por meio de um estudo acerca de como operam os juristas, mas pela reprodução do discurso interno da própria comunidade dos juristas. Infelizmente, descobrimos nesta pandemia que muitos médicos se comportam exatamente dessa forma, confiando mais m sua intuição e nos cânones que regem a sua prática cotidiana do que nas pesquisas empíricas mais atualizadas.

Na cultura jurídica contemporânea, o conhecimento dos fatos não desempenha um papel central, tanto que o ensino jurídico se concentra em garantir que o repertório cultural dos estudantes deve ser equacionado ao repertório cultural dos profissionais, o que permite que o jurista em formação aprenda a julgar adequadamente como um certo argumento deverá ser recebido pela comunidade dos juristas. Por mais que não haja garantias no sentido de que os

processos serão julgados pelos mesmos critérios de aceitabilidade, essa geração de um discurso padrão é uma forma socialmente eficaz de dar certa organização ao direito.

Essa é a estrutura da dogmática e ela garante que o conhecimento tradicional (com seus erros e acertos, com seus limites e possibilidades) seja reproduzido pelas gerações seguintes. Existe, claro, uma certa variação, pois os sentidos dominantes vão sendo transformados. Mas essa variação segue a lógica consuetudinária dos discursos dominantes: pequenas variações se somam ao longo do tempo, gerando trânsitos maiores, como ocorre na língua falada e escrita. Esse é um processo que produz heterodoxias discretas, das quais algumas são normalizadas pela prática: o princípio da proporcionalidade, a ponderação de princípios, a interpretação teleológica, o uso da jurisprudência como argumento central.

Essa já foi a estrutura dos conhecimentos da medicina tradicional e da alquimia, e continua sendo a estrutura da astrologia, da acupuntura, da teologia e do direito: existe um repertório de saberes compartilhados e a adequação de um diagnóstico e de uma prescrição são medidos em termos dedutivos: as teses defendidas são deduções adequadas dos princípios aceitos ou não?

Não existe uma pesquisa astrológica, uma pesquisa tarológica, uma pesquisa teológica. Existem textos fundantes, existem interpretações ortodoxas, existem livros que noticiam as discussões dos sábios. Temos a intuição de que esses conhecimentos decorrem de uma longa decantação de estratégias hermenêuticas, da criação de categorias que têm oferecido respostas interessantes e têm contribuído para que as pessoas desenvolvam um "autoconhecimento", no sentido de terem narrativas sobre a própria subjetividade. Trata-se de um saber construído ao longo do tempo, a partir da acumulação de experiências,

Em contrapartida, existem pesquisas biológicas, sociológicas e psicológicas, feitas por pessoas que coletam dados acerca de certos fenômenos, em busca de compreender como eles ocorrem no

mundo. Essa abertura para os dados é a marca da ciência: a busca de construir um discurso baseado em evidências, em informações obtidas da observação direta do mundo.

3. Entre ciência e arte

A religião não é um saber científico. Tampouco é científico o senso comum que nos dá a maior parte de nossas percepções de como o mundo é, de como as coisas operam. A moralidade pode ser um saber sobre o que é certo e errado, mas não é uma ciência. As artes não são saberes, mas são competências que articulam saberes e habilidades.

Os juristas atuais estão mais próximos dos artistas do que dos cientistas. Eles conhecem o seu público e por isso sabem que certas intervenções são capazes de mobilizar sentimentos, de estimular ações, de desencadear apreciações positivas e negativas. Tal como ocorre com atores ou músicos, o fato de os juristas compartilharem com seu público uma mesma sensibilidade faz com que eles se tornem capazes de mobilizar a atenção de seu auditório.

A educação dos advogados faz com que eles partilhem, em grande medida, a visão de mundo dominante entre os juízes, o que possibilita que eles mobilizem uns aos outros por meio da construção de discursos que refletem suas intuições e subjetividades. Se não houvesse essa identidade, a elaboração de abordagens retóricas eficientes exigiria um raciocínio estratégico calcado em um conhecimento meticuloso do auditório. Esse tipo de abordagem externa (que envolve olhar a cultura jurídica de fora) exige muito esforço, muito cálculo e muita informação. Já as abordagens internas operam por meio de uma identidade, e não de um cálculo, o que possibilita que estratégias retóricas eficientes sejam alcançados com mais intuição e menos conhecimento.

Pode chegar um momento em que o artista venha a ser substituído

pelo cientista. É possível que cheguemos a um momento em que a análise psicológica nos mostre tanto sobre o modo como a música desencadeia reações orgânicas, e sobre quais são as músicas que as pessoas preferem, que um algoritmo que operacionalize esses conhecimentos seja mais eficiente, enquanto músico, do que um artista com uma intuição bem cultivada.

Há algumas décadas, imaginava-se que nenhuma máquina seria capaz de ganhar dos homens em jogos complexos, como o xadrez. Assim como hoje muitos imaginam que uma petição inicial nunca poderá ser feita (ou apreciada) com a mesma qualidade por humanos e computadores. E mais pessoas ainda imaginam que um julgamento operado por algoritmos tende a perder uma dimensão de humanidade, que seria fundamental na realização da justiça.

Como aponta Kahneman, haverá um ponto no futuro em que os computadores se tornarão capazes de gerar soluções mais complexas, mais eficientes e inclusive mais justas do que aquelas alcançadas por seres humanos. Mas nós ainda estamos longe desse ponto, que não sabemos quando chegará. Exercícios de futurologia são instigantes, mas não são os melhores guias para as nossas ações mais imediatas.

De um modo ou de outro, estamos chegando próximos ao momento em que conhecimentos empíricos sólidos serão guias mais adequados para a ação do que intuições baseadas em uma subjetividade compartilhada. Para isso, não precisamos de algoritmos avançados de inteligência artificial, mas apenas de um acesso mais amplo aos dados e de uma análise cuidadosa das informações disponíveis. Não falamos aqui do momento em que a inteligência humana será superada pela inteligência artificial, mas do momento em que o conhecimento empírico oferecerá uma base mais sólida do que a intuição bem treinada de um técnico experiente. Esse é um limiar que a medicina atingiu no século XX e que parece provável que o Direito atinja em um prazo relativamente curto.

Para que o trabalho metódico do cientista do direito ofereça bases

mais seguras para a ação do que a intuição de um jurista experiente, ele precisa de muitos dados. A grande vantagem da dogmática é que uma pessoa pode executar uma prática competente a partir de dados muito lacunares, pois ela trabalha com estereótipos, simplificações e generalizações. A dogmática jurídica nos permite escrever uma petição inicial em uma ação de indenização por responsabilidade civil, usando o mínimo necessário de dados para promover uma decisão favorável. O que importa não é a riqueza das informações, mas a capacidade de reduzir as informações úteis às categorias da dogmática, que serão as únicas (espera-se...) utilizadas no processamento humano do caso.

Mas chegará um ponto em que teremos dados suficientemente amplos sobre as decisões judiciais, sobre os juízes e sobre os casos. Nesse momento, pode ser que um algoritmo bem montado seja capaz de nos dizer mais sobre as perspectivas de sucesso de uma ação (e sobre as estratégias jurídicas adequadas) do que os saberes de um jurista experiente. Esse é um trânsito que não ocorre do dia para a noite. Precisamos de dados, e as pesquisas são máquinas de produzir dados, seja pela observação da realidade ou pela produção artificial de experimentos.

A peculiaridade da pesquisa é que ela produz dados que se encaixem em metodologias que permitem produzir informações a partir deles: que nos permite falar de entidades mais abstratas, a partir de informações coletadas acerca de unidades mais concretas.

Sem dados, não há ciência. Mas os dados não geram a ciência imediatamente: eles precisam ser classificados segundo taxonomias definidas, as quais são desenvolvidas por abordagens teóricas. Somente esse tipo de classificação permite que os dados brutos sejam interpretados, que eles sejam usados como evidências para sustentar determinadas descrições ou explicações da realidade.

No direito, obter dados era uma atividade muito penosa, até que o processo de informatização dos tribunais converteu dados que eram muito difíceis de acessar em dados potencialmente acessíveis. Esse

processo de informatização produziu uma multiplicidade de informações que não foram suficientemente analisadas, o que deslocou o desafio de compreender o direito: não tempos mais uma carência de dados, mas uma carência de interpretações sobre eles.

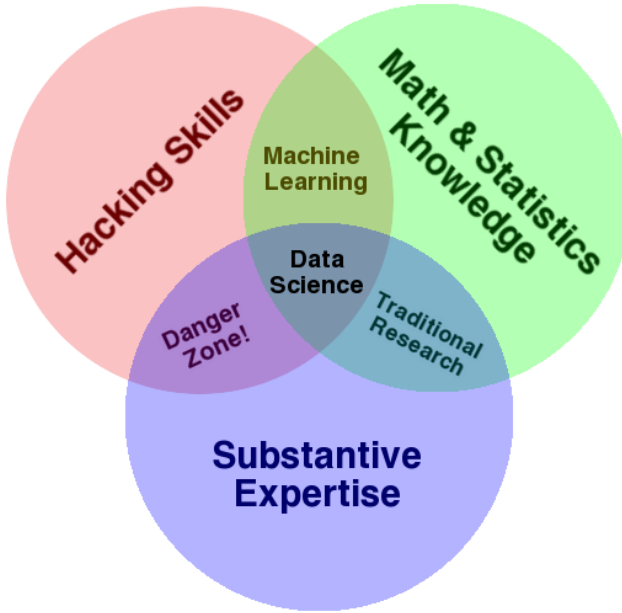
Essa carência é determinada pelo fato de que os juristas não foram treinados para lidar com os fatos do mundo, mas apenas para produzir discursos dogmáticos sólidos. Há pessoas suficientes para analisar os dados, eles são suficientemente disponíveis, existem recursos computacionais capazes de nos permitir a realização de várias formas de processamento desses dados. Continua sendo relevante produzir dados (especialmente dados que não foram bem mapeados), mas tem ganhado relevância o fato de que temos uma quantidade imensa de dados que não foram tratados, classificados e analisados.

E nem falamos aqui de big data, mas apenas de dados comuns: decisões judiciais, dados processuais, tempos de julgamento, argumentos típicos e atípicos. Para que desenvolvamos modelos descritivos e explicativos suficientemente robustos para compreender os padrões pelos quais a atividade jurídica se desenvolve, existe um longo caminho a seguir, especialmente no plano teórico: não temos classificações adequadas para essa função. As taxonomias jurídicas são voltadas para a dogmática, para a reprodução dos saberes tradicionais, e não para o desenvolvimento da pesquisa.

4. Pesquisa de Dados

Ciência de Dados é um rótulo um pouco vago, que apesar das suas limitações, "is perhaps the best label we have for the cross-disciplinary set of skills that are becoming increasingly important in many applications across industry and academia" (VanderPlas, 2017). Uma das descrições mais interessantes deste campo é o seguinte diagrama, elaborado por Drew Conway (2015[2]):

[2] http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram



Do ponto de vista da pesquisa acadêmica, que Conway coloca na interseção entre o conhecimento material (incluindo o metodológico) e o conhecimento estatístico, a principal inovação é a utilização de habilidades computacionais que servem como mecanismos de coleta, organização e análise dos dados. Essa abordagem nos permite falar em pesquisa de dados: não nos interessa definir se existe propriamente uma ciência de dados, mas interessa-nos a capacidade de incorporar como insumos de nossas pesquisas os dados que se tornaram acessíveis a quem tem as habilidades computacionais para acessá-los e organizá-los de forma a que possam ser analisados segundo os parâmetros da pesquisa científica.

Abordar o direito a partir da **data science** envolve enfrentar o desafio de como as práticas jurídicas podem ser explicadas a partir de interpretações realizadas a partir de bases de dados. Essas

interpretações consistem na busca de padrões, de algum tipo de regularidade que nos permita utilizar um conjunto informações particulares (sobre processos, sobre decisões, sobre ministros), para fazer afirmações gerais sobre o conjunto de dados.

Para enfrentar esse desafio, o primeiro passo é aprender como é possível fazer pesquisa a partir de bases de dados já existentes, desenhando pesquisas capazes de construir novos conhecimentos, a partir de informações previamente organizadas.

Porém, nem sempre as bases disponíveis são suficientes para enfrentar os nossos problemas de pesquisa, o que pode exigir a construção de novas bases ou, no mínimo, a complementação de bases existentes. Essa complementação normalmente se dá por meio da criação e implementação de novas **classificações**, que permitam segmentar os dados segundo parâmetros diversos dos que vinham sendo utilizados.

O desafio geral é encontrar padrões, mas os padrões somente são formados quando classificamos os dados de uma maneira determinada. A **classificação** é o grande desafio teórico e filosófico envolvido na pesquisa, pois ela envolve o desenvolvimento de modelos conceituais capazes de captar as complexidades dos objetos que pretendemos descrever. Sem uma teoria adequada, é impossível fazer pesquisa empírica com resultados sólidos e é muito fácil chegar a conclusões equivocadas, ainda mais quando se utiliza modelos de machine learning.

Os modelos de machine learning são algoritmos desenvolvidos para buscar padrões e eles encontram padrões em quase qualquer conjunto de informações. Ferramentas de clusterização, por exemplo, vão subdividir um conjunto em subconjuntos, a partir de critérios de semelhança. Ocorre que esses modelos encontram **padrões nas informações**, não encontram **padrões nos fatos**. Para que eles possam ser úteis, é preciso converter os fenômenos observados empiricamente em informações com sentido, o que exige conhecimento material profundo dos objetos analisados. Sem um

modelo descritivo adequado, não é possível aplicar as ferramentas computacionais disponíveis.

Portanto, a observação de padrões significativos em um conjunto de dados exige a combinação de **conhecimento material** (que garanta classificações adequadas) com habilidades para identificar padrões, que são oferecidas pelo conhecimento matemático e estatístico. Sem uma estrutura teórica robusta, a aplicação de estratégias matemáticas pode gerar conclusões extremamente problemáticas.

Embora o diagrama acima qualifique justificadamente como Danger Zone! a combinação de conhecimentos materiais e computacionais, sem habilidades estatísticas capazes de justificar as inferências, existe uma outra zona de risco muito grave: a existência de conhecimentos teóricos limitados, que conduz ao manuseio de categorias teóricas incapazes de organizar devidamente a experiência, gerando descrições distorcidas e explicações falsas.

Para completar o tripé da data science, precisamos desenvolver **habilidades computacionais** que viabilizem a busca e o tratamento das informações disponíveis, com a construção de algoritmos de coleta e análise dos dados.

Saindo do diagrama acima, consideramos que é preciso adicionar um outro campo fundamental para o uso acadêmico da "data science": a habilidade de desenhar estratégias de pesquisas capazes de gerar conclusões confiáveis a partir dos dados existentes, e que é tipicamente chamada de **metodologia**. Que estratégias metodológicas são viáveis para que seja possível compreender o modo como alguns fenômenos observáveis se relacionam? Estratégias baseadas em matemática e estatística são apenas uma parte desse jogo, mas não esgotam o repertório de conhecimentos necessários.

Embora falemos em termos de ciência de dados, devemos admitir que se trata de uma concessão ao fato de que essa tem sido uma categoria amplamente difundida e que atrai a atenção das pessoas. Independentemente desses rótulos, entendemos que o processo de informatização alterou a própria estrutura do direito (mudando

a forma como os dados, especialmente as informações judiciais, são produzidos e utilizados pelos juristas) e que também mudou radicalmente as habilidades necessárias para que um jurista opere dentro do contexto contemporâneo.

Essas alterações na estrutura do direito, nas formas de conhecê-lo e nas atividades práticas dos juristas exige uma combinação nova de habilidades: não basta um conhecimento material sobre os cânones dogmáticos, mas é preciso desenvolver também competências metodológicas, estatísticas e computacionais, que se tornaram essenciais para que o jurista atue como cientista (identificando padrões em um conjunto amplo de dados) e não apenas como artista (modelando discursos a partir de sua própria subjetividade).