DOI: 10.5433/1981-8920.2016v21n2p307

A FORMAÇÃO EM CIÊNCIA DE DADOS: UMA ANÁLISE PRELIMINAR DO PANORAMA ESTADUNIDENSE

FORMACIÓN EN CIENCIAS DE DATOS: UN ANÁLISIS PRELIMINAR DE LAS PERSPECTIVAS DE EE.UU.

Renata Gonçalves Curty*
Jucenir da Silva Serafim**

RESUMO
Introdução: Cientistas de dados têm recebido grande destaque nos últimos anos seguindo as demandas do mundo do trabalho estimuladas pela ciência aberta e pela era *big data.* Amplamente divulgada em 2008, e agora presente nos mais diferentes setores e aplicações, a terminologia "cientista de dados" foi anunciada em 2012 como a mais atraente e uma das mais bem remuneradas do século XXI, culminando em uma crescente oferta de cursos de formação.
Objetivo: Caracterizar e compreender os aspectos formativos do cientista de dados.
Metodologia: O artigo relata um recorte de uma pesquisa de levantamento com base na análise preliminar de 93 cursos em ciência de dados ofertados por instituições estadunidenses.
Resultados: A análise de conteúdo das informações contidas nos *websites* dos programas identificados permitiu evidenciar que este profissional é formado para lidar com aspectos relacionados à coleta, tratamento, transformação, análise, visualização e curadoria de grandes e heterogêneas coleções de dados orientadas à resolução de problemas práticos e reais.
Conclusão: Foi possível constatar que, de modo geral, a formação em ciência de dados atribui grande ênfase a habilidades estatísticas, matemáticas e computacionais, incluindo programação e modelagem avançada, sendo que muitas destas são pré-requisitos para ingresso nestes cursos.

Palavras-chave: Ciência de Dados. Cientista de Dados. Competências Profissionais. Formação Profissional.

*Philosophy Doctor (PhD) e Master in Philosophy (MPhil) em Information Science and Technology pela School of Information Studies (Syracuse University, NY). Professora do Departamento de Ciência da Informação da UEL. E mail: renatacurty@uel.br
**Mestrando do Programa de Pós-graduação em Ciência da Informação da Universidade Estadual de Londrina

1 INTRODUÇÃO

As organizações e a ciência de modo geral dependem diretamente de dados. É notório o papel de dados enquanto insumos essenciais para o processo de tomada de decisão, bem como para o avanço científico. Porém, apenas recentemente a ciência de dados tem se institucionalizado como educação formal, e os profissionais com esta formação têm ocupado posição de destaque no mundo do trabalho.

Listada como uma das profissões mais relevantes até 2020 pelo Fórum Econômico Mundial e anunciada como a profissão mais "*sexy*" do século XXI pela *Harvard Business Review*, a posição de cientista de dados teve sua terminologia cunhada por Patil e Hammerbacher em 2008. Posteriormente, este termo passou a ser adotado pelo *LinkedIn* e pelo *Facebook* em anúncios de vagas buscando profissionais para lidar com o grande volume e tráfego de dados nas mídias sociais (DAVENPORT; PATIL, 2012).

O recrutamento de cientistas de dados tem sido cada vez mais recorrente nas postagens de posições e oportunidades de emprego. Uma busca recente[1] na plataforma do *LinkedIn* por vagas com a designação "*data scientist*" resultou em cerca de 2.500 ofertas vigentes de emprego nos EUA, em universidades e na indústria, nos mais variados setores de aplicação, incluindo saúde, comércio eletrônico, setor bancário, consultoria de investimentos e negócios, ramo imobiliário, entre outros.

Outra fonte que atesta o crescente interesse mundial por essa área é o *Google Trends*[2] que demonstra um ápice, a partir de 2012, de buscas submetidas ao Google por usuários de diferentes países (Índia, Cingapura, Estados Unidos, Austrália e Canadá), utilizando o termo "*data scientist*" e "*data science*", geralmente combinadas com termos como formação, cursos, salário, habilidades e certificação profissional.

[1] Dados atualizados em 25/10/2016.
[2] https://www.google.com/trends

Não obstante, a demanda por profissionais com capacidades analíticas e técnicas para lidar com grandes e heterogêneos volumes de dados, tem resultado em uma expansão na oferta de cursos para a formação deste perfil profissional. Todavia, pouco é sabido e documentado na literatura científica nacional e também internacional a respeito destes cursos de formação, bem como sobre as habilidades e competências esperadas dos egressos. Desse modo, este artigo descreve os resultados preliminares de um levantamento feito nos *websites* de diferentes cursos em ciência de dados ofertados por universidades e escolas estadunidenses de modo a responder aos seguintes questionamentos: quais são as principais características desses cursos? Quais pré-requisitos são esperados dos ingressantes desses cursos? Quais objetivos e competências formativas gerais esses programas se propõem a abordar?

As respostas aos questionamentos supracitados visam não somente contribuir para uma visão panorâmica da educação voltada à ciência de dados no território estadunidense, mas também fornecer subsídios para a discussão acerca da formação do cientista de dados no Brasil, levando em conta a demanda gerada com o avanço da ciência orientada a dados e pela aclamada era *big data*. Para tanto, em um primeiro momento serão apresentados aspectos da formação e atuação do cientista de dados com base na literatura vigente. Posteriormente, será descrito o percurso metodológico adotado pela pesquisa para obtenção dos dados empíricos, seguido da análise e discussão dos resultados. Por fim, serão tecidas as considerações finais, incluindo reflexões sobre as implicações dos achados da pesquisa, bem como os direcionamentos futuros deste estudo.

2 **FORMAÇÃO E ATUAÇÃO DO CIENTISTA DE DADOS:** INDÍCIOS NA LITERATURA

Nos últimos anos, experimentamos um crescente acúmulo de dados disponíveis na grande rede. Na chamada era *big data* quatro “Vs” são imperativos: o volume (quantidade de dados), a variedade (heterogeneidade e diversidade de formatos e tipos de dados estruturados e não-estruturados) e a velocidade (produção de dados em fluxo ininterrupto e em tempo real) (LANEY,

2001). Há também os que acrescentam outros dois "Vs": a viabilidade e o valor. A viabilidade se refere à condição de identificar relacionamentos entre variáveis e padrões latentes em grandes quantidades de dados, enquanto o valor diz respeito à necessidade de aplicação e tradução desses relacionamentos e padrões a situações reais e práticas que tragam resultados tangíveis (BIEHN, 2013).

Ao realizarem uma metanálise de diferentes definições de *big data*, dentre as quais algumas que se apropriam dos "Vs", Ward e Barker (2013) concluem que existem três fatores críticos comuns carregados nesses conceitos: o tamanho, ou seja o grande volume dos conjuntos de dados, a complexidade, que corresponde à estrutura, comportamento e permutações dos conjuntos de dados, e, por fim, as tecnologias, ferramentas e técnicas que são utilizadas para processar um conjunto de dados bastante grande e complexo.

Dados gerados em tempo real e em fluxo contínuo provenientes de *logs* de sistemas, sensores, satélites, redes sociais, registros de transações *online*, dados brutos e primários de pesquisa resultantes de estudos financiados, pesquisas cidadãs e coletivas, dados abertos governamentais, entre outros, têm se acumulado exponencialmente em servidores *web*. O grande desafio, portanto, não é o de encontrar ou localizar dados, mas sim descobrir o que fazer com eles e como utilizá-los de modo significativo, e utilizando o seu máximo potencial de aplicação. Há, desse modo, uma crescente demanda de análise de dados por meio de uma abordagem holística e interdisciplinar, que considere a integração e a combinação de dados provenientes de diferentes fontes, e é justamente esta abordagem que define a ciência de dados (LOUKIDES, 2012).

Finzer (2013) relata que o termo ciência de dados foi mencionado pela primeira vez em 2001, em um texto de autoria de William S. Cleveland intitulado, em tradução livre, "Ciência de dados: um plano estratégico para a

expansão das áreas técnicas no campo da Estatística"[3], o qual tinha por objetivo aliar a estatística à programação e à computação. No entanto, Cleveland (2001) utiliza a nomenclatura *data analyst* e pouco descreve as características deste profissional.

O relatório britânico encomendado pela *Joint Information Systems Committee* (JISC) acerca das habilidades, dos papéis e da carreira dos cientistas de dados, reconhece a dificuldade de um consenso quanto à definição deste profissional, mas o define de modo geral como aquele que trabalha *in loco* onde pesquisas são realizadas, ou em estreita colaboração com os pesquisadores ou grupos de cientistas em centros de dados, e que está envolvido na investigação criativa e de análise de dados, oferecendo soluções tecnológicas para o manuseio e uso de dados digitais (SWAN; BROWN, 2008). Entretanto, é importante ponderar que esta descrição é um tanto restritiva em comparação a outras encontradas na literatura, pois as autoras definem outros papéis como gerenciadores de dados (*data manager*) e bibliotecários de dados (*data librarians*), subdividindo suas atribuições na ecologia da ciência de dados.

A despeito da dificuldade de se caracterizar a profissão de modo uníssono, a literatura é unânime em enfatizar que os cientistas de dados devem apresentar domínio estatístico e computacional para a programação e uso de sistemas capazes de processar grandes volumes de dados (CHATIFELD *et al*., 2014; GRANVILLE, 2014) e ter capacidade de explorar a inteligibilidade em dados que a princípio são desestruturados e sem sentido (VAN DER AALST, 2014). Granville (2014) descreve o cientista de dados como um profissional generalista que conhece negócios, estatística, ciência da computação, capaz de relacionar alguns conhecimentos específicos entre os quais arquitetura de dados, comunicação no ambiente empresarial e outros.

Conaway (2010) acrescenta mais uma dimensão de habilidades, ao articular que a ciência de dados constitui-se como a congruência entre domínios computacionais, habilidades de matemática e estatística, e

[3] Do original: "*Data Science: an action plan for expanding the technical areas of the field of Statistics*".

especialidade na área de aplicação de dados. Nesse sentido, Finzer (2013) detalha que em primeiro lugar, há o pensamento disciplinado quantitativo encontrado na matemática e na estatística. A partir da estatística vem a compreensão da variabilidade e experiência no uso de ferramentas de análise para o trabalho com dados. Em segundo lugar, há a necessidade de experiência substantiva a qual confere ao cientista de dados uma compreensão do contexto disciplinar sem a qual a escolha de uma metodologia válida de análise será difícil ou impossível. Em terceiro lugar, existem as habilidades de computação e de análise de dados que, combinadas com a criatividade e com a capacidade de resolução de problemas, permitem aos cientistas de dados visualizar a estrutura dos dados e extrair sentido dos mesmos (FINZER, 2013, tradução nossa).

Para Stanton *et al.* (2012) a formação do cientista de dados é fortemente interdisciplinar. Nesse aspecto, os autores complementam que os cientistas de dados são responsáveis pela identificação, coleta, tratamento, transformação, análise, visualização e curadoria de grandes conjuntos de dados heterogêneos. Muito embora o aspecto da análise tenha uma ênfase de atribuição significativa – o que explica em parte o uso intercambiável dos termos "*data science*" e "*data analytics*" –, os cientistas de dados também devem ter um profundo entendimento de como dados foram coletados/produzidos, pré-processados e transformados. Estes processos influenciam diretamente na seleção de métodos e ferramentas analíticas mais apropriadas, além de como os resultados provenientes destes métodos devem ser interpretados e comunicados.

Além de possuir uma amplitude de experiência nas áreas de curadoria, análise, ciberinfraestrutura e o necessário domínio na área de aplicação dos dados, o que diferencia o cientista de dados de outras especialidades profissionais é a ênfase no atendimento às necessidades de dados de usuários e tomadores de decisão (STANTON *et al.*, 2012).

Para superar o discurso especulativo ou idealizador sobre a profissão, Kim e Lee (2016) realizaram a análise de cerca de mil ofertas de emprego para cargos de cientistas de dados de 736 empresas registradas em três diferentes

plataformas *online* de recrutamento dos EUA: Monster.com, Indeed.com e CareerBuilder.com.

Com esse estudo, os autores puderam identificar três grandes classes de habilidades e conhecimentos profissionais essenciais do ponto de vista dos empregadores: sistemas, negócios e técnicas. A habilidade em sistemas é subdividida em duas subclasses: gerenciamento e soluções de problemas. A de negócios compreende as subclasses: social, de negócios propriamente e gerencial. Finalmente, a técnica abarca as subclasses: *software*, arquitetura e redes e *hardware*. O detalhamento destas classes e subclasses, com a listagem de habilidades por ordem decrescente de frequência está descrito na Tabela 1.

Tabela 1 - Habilidades e Conhecimentos Essenciais para os Cientistas de Dados

<table>
<tr><th>Sistemas</th><th>N.</th><th>%</th><th>Negócios</th><th>N.</th><th>%</th><th>Técnicas</th><th>N.</th><th>%</th></tr>
<tr><td>Desenvolvimento</td><td>1230</td><td>99</td><td>Social</td><td>1148</td><td>93</td><td>Software</td><td>1189</td><td>96</td></tr>
<tr><td colspan="3">● Análise
● Implementação/Teste
● Projeto
● Gestão de dados
● Conhecimento de diferentes tecnologias
● Desenvolvimento de Metodologias
● Programação
● Operações/Manutenção
● Integração
● Documentação</td><td colspan="3">● Habilidades Interpessoais
● Comunicação
● Auto-motivação</td><td colspan="3">● Linguagem de programação
● Banco de Dados/Data Warehouse
● Plataformas Open Source Domínio de diferentes pacotes
● Visualização de dados</td></tr>
<tr><td>Solução de Problemas</td><td>1227</td><td>99</td><td>Negócios</td><td>1130</td><td>91</td><td>Arquitetura e Redes</td><td>714</td><td>58</td></tr>
<tr><td colspan="3" rowspan="2">● Modelagem de dados
● Análise quantitativa/estatística
● Pensamento analítico/lógico
● Criatividade/inovação
● Capacidade para solução de problemas
● Adaptabilidade/flexibilidade
● Capacidade estratégica</td><td colspan="3">● Conhecimento específico do setor/negócio
● Habilidade de análise macro
● Negócios online/e-commerce</td><td colspan="3">● Internet/LAN/WAN
● Networking e dispositivos de rede
● Computação nas nuvens
● Computação cliente-servidor
● Arquitetura e segurança de rede
● Computação ubíqua
● Sistemas legados/mainframes</td></tr>
<tr><td>Gerencial</td><td>1019</td><td>82</td><td>Hardware</td><td>442</td><td>36</td></tr>
</table>

	• Administração Geral • Organização/Liderança • Capacidade de monitoramento e controle • Planejamento • Treinamento • Gestão de mudança • Gerenciamento de projetos	• Dispositivos de Armazenamento • Impressoras • Desktop/PC • Servidores • Estações de Trabalho (Workstation) • Conhecimentos gerais de hardware

Fonte: Elaborado com base em Kim e Lee (2016, p. 166, tradução nossa).

Em seu último relatório sobre ciência de dados, a empresa de consultoria *CrowdFlower* identificou, por questionamento direto a profissionais que ocupam cargos como cientistas de dados, que entre as atividades mais desempenhadas por estes profissionais estão o tratamento e organização de dados, a exploração de correlações e de padrões no conjunto de dados e o refinamento de algoritmos (CROWDFLOWER, 2016). Adotando abordagem semelhante à adotada por Kim e Lee (2016) para identificação de habilidades e competências dos cientistas de dados, o mesmo relatório realizou análise da descrição de vagas de cerca de quatro mil postagens no *LinkedIn*, e permitiu identificar quais ferramentas analíticas, sistemas e linguagens de programação são dominantes nos quesitos descritos em vagas de emprego para cientistas de dados divulgados por meio desta plataforma *web* (Tabela 2).

Tabela 2 - As dez ferramentas/sistemas essenciais aos cientistas de dados

Ferramenta /Sistema	Descrição	Vagas com Menções	% de Menções
SQL	Linguagem de Consulta Estruturada para pesquisa declarativa padrão para banco de dados relacional	1987	56%
Hadoop	Plataforma de software em Java de computação distribuída voltada para clusters e processamento de grandes massas de dados	1713	49%
Python	Linguagem de programação de alto nível, interpretada, de script, imperativa, orientada a objetos, funcional, de	1367	39%

	tipagem dinâmica e forte, cuja principal característica é de simplificar o esforço de programação (redução das linhas de código)		
Java	Linguagem de programação interpretada orientada a objetos	1287	36%
R	Ambiente de desenvolvimento integrado para cálculos estatísticos e de gráficos, com avançados recursos de programação em linguagem de igual nomenclatura	1120	32%
Apache Hive	Uma infraestrutura de armazém de dados para a compactação, consulta e análise de dados	1099	31%
Mapreduce	Modelo de programação, e framework introduzido pelo Google para suportar computações paralelas em grandes coleções de dados em clusters de computadores	768	22%
NoSQL	Classe de bancos de dados não-relacionais	657	18%
Apache Pig	Plataforma para análise de grandes conjuntos de dados que consiste em uma linguagem de alto nível para expressar programas de análise de dados, juntamente com a infraestrutura para a avaliação desses programas	561	16%
SAS	Sistema integrado de aplicações para o processamento e análise estatística de dados, consistindo em módulos de acesso e recuperação de dados, gerenciamento de arquivos, rotinas de geração de gráficos e geração de relatórios	560	16%

Fonte: Elaborado com base em CrowdFlower (2016, p. 9, tradução nossa).

Cumpre frisar que os dados de frequência e importância encontrados em Kim e Lee (2016), quanto ao domínio técnico de ferramentas e sistemas específicos, divergem em alguns momentos dos acima descritos, até mesmo pelo fato de os estudos terem avaliado quantidades de vagas diferentes, provenientes de diferentes plataformas. Ainda assim, todos os sistemas e ferramentas presentes na Tabela 2 foram elencados nos resultados de

pesquisa de Kim e Lee (2016), fator esse que ajuda a confirmar a relevância destes para a atuação profissional dos cientistas de dados.

Como visto, estudos anteriores empreenderam esforços para identificar as demandas do mundo do trabalho com relação aos cientistas de dados, por meio de pesquisas diretas com estes profissionais, e até mesmo por meio de análises de anúncios de ofertas de emprego. Muito embora a profissão ainda esteja em fase de consolidação, há um notável progresso da comunidade europeia e da América do Norte em capacitar e qualificar cientistas de dados em nível de pós-graduação, por meio de treinamento informático, estatístico e de conteúdo direcionado a essa especialidade de formação (SWAN; BROWN, 2008). Este avanço tem sido expresso na crescente oferta de cursos em massa *online* e abertos os chamados *Massive Open Online Courses* (MOOCs) e cursos pagos na modalidade *online* e presencial para o desenvolvimento das habilidades e competências em ciência de dados.

O presente artigo analisou, seguindo a abordagem metodológica descrita a seguir, a composição destes diferentes cursos oferecidos nos Estados Unidos com o objetivo de traçar um panorama geral sobre a formação de cientistas de dados.

3 PROCEDIMENTOS METODOLÓGICOS

Para a identificação preliminar de cursos relacionados à ciência de dados, esta pesquisa utilizou a compilação feita pelo *website* <www.masterindatascience.org>, conhecido por fornecer informações sobre ciência de dados para os que consideram uma carreira nesta área. A coleta de dados transcorreu entre os meses de agosto e setembro de 2016, e teve como critério inicial incluir todos os cursos de pós-graduação e certificações em estudos relacionados à ciência de dados nos EUA. A delimitação para programas estadunidenses justificou-se pelo fato de a formação e a demanda por esta categoria profissional ter origem nesse país, bem como a conveniência, por já existirem iniciativas de identificação destes programas, o que seria inviável de ser levantado por meio de mecanismos de buscas e

pesquisas sistemáticas em fontes dispersas, com a mesma agilidade e relevância.

O procedimento de coleta de dados consistiu no acesso, de modo individual, ao *website* de cada um dos programas listados no diretório de cursos supracitado. Isto porque o *website* listava apenas informações cadastrais dos programas, não provendo os detalhamentos necessários à pesquisa.

A pesquisa de levantamento se atentou especificamente à identificação dos seguintes itens: instituição, localidade (estado), nome do programa, data de implementação, modalidade de oferta, carga horária, tempo total de duração, número de créditos exigidos, pré-requisitos, objetivos, descrição curricular, data de coleta e página *web* do programa. Todas as informações coletadas foram registradas em planilha Excel para organização e tratamento dos dados.

Num primeiro momento foi identificado um total de 410 diferentes cursos e certificações relacionados à ciência de dados, ofertados por universidades e instituições de ensino distribuídas em 41 dos 50 estados dos EUA e no Distrito Federal. No entanto, após uma pré-avaliação das diferentes titulações concedidas e das especificações dos cursos, observamos certa dispersão temática ou abordagem demasiadamente periférica a questões concernentes ao escopo da ciência de dados. Muitos cursos dos cursos identificados tinham nomenclaturas, tais como: Sistemas de Informação (*Information Systems*), Informática na Saúde (*Health Informatics*), Gerenciamento de Operações (*Operations Management*), entre outros. Portanto, para evitar o desvio dos objetivos originais propostos pela pesquisa, decidimos filtrar os resultados, e manter para fins de análise final apenas a parcela de cursos e certificações que incluíam em seus nomes os termos "*data science*" e/ou "*data analytics*". Embora este último seja mais restritivo à questão analítica, sua inclusão foi motivada pelo frequente uso intercambiável desses dois termos (STANTON *et al.*, 2012).

Utilizamos diferentes ferramentas de apoio para fins de análise dos dados da amostra selecionada. O Excel foi utilizado para computar

quantificações simples, como média, quantidades e percentuais para itens de dados mais objetivos e possíveis de uniformização, incluindo: universidades, estados, números de créditos e carga-horária. Também utilizamos o recurso Google *MyMaps*[4], para representar a localização geográfica dos programas que fizeram parte da amostra e para demonstrar por meio de visualização de dados a distribuição de cursos no território dos EUA. Para a análise do *corpus* textual, por meio da frequência de ocorrências de termos derivados da nomenclatura da titulação/curso, objetivos e pré-requisitos exigidos, utilizamos como ferramenta de apoio o sistema *Online NgramAnalyzer*[5], em alguns momentos combinado com o sistema *TagCrowd*[6] para a visualização destas ocorrências.

4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Nas seções subsequentes serão apresentados e discutidos os resultados da análise de 93 programas selecionados que continham em sua nomenclatura os termos "*data science*" e/ou "*data analytics*" (APÊNDICE A), com relação aos seguintes aspectos: características gerais, objetivos e competências formativas gerais, e pré-requisitos para o ingresso.

3.1 Das Características Gerais

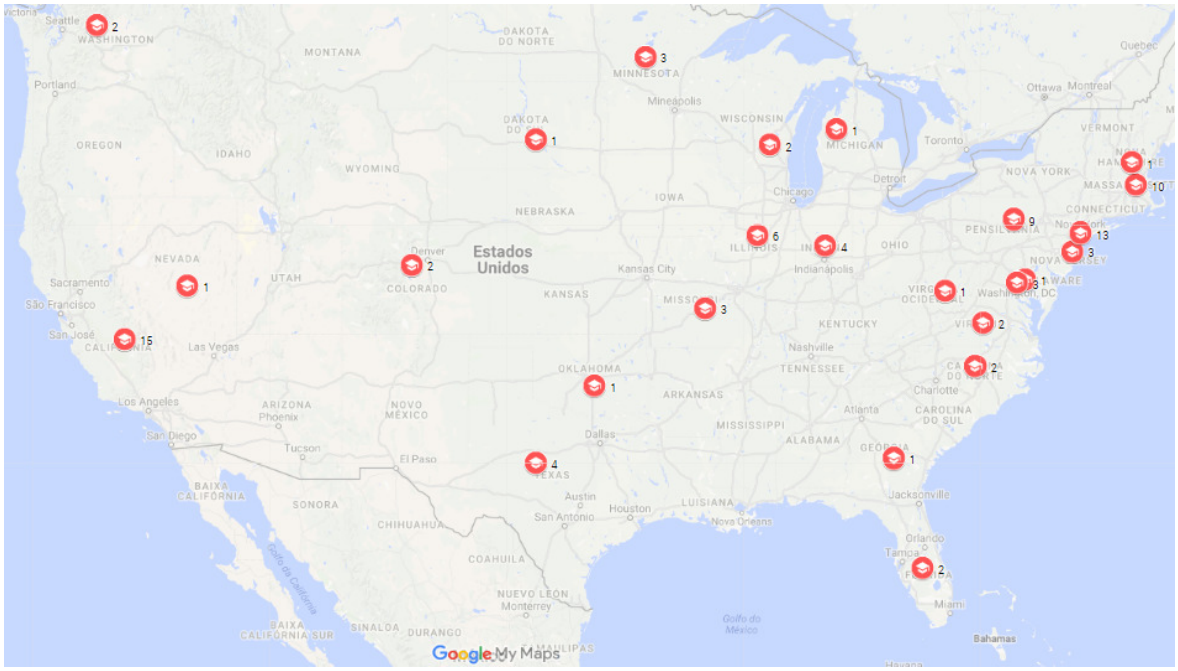
Os 93 programas em ciência de dados que fizeram parte da amostra estão distribuídos em 24 estados e no distrito federal, conforme ilustrado na Figura 1.

[4] http://www.google.com/mymaps

[5] sistema *online* e gratuito de apoio à mineração de dados, o qual permite quantificar a contiguidade sequencial de palavras, termos compostos de duas até cinco palavras, bem como classificar, aglutinar e comparar dados textuais estruturados e não-estruturados. http://guidetodatamining.com/ngramAnalyzer/

[6] http://tagcrowd.com/

Figura 1 - Presença de Cursos em *Data Science* e/ou *Data Analytics* nos EUA.



Fonte: Visualização gerada a partir de dados da pesquisa.

Embora a presença de cursos na faixa leste do país esteja em maior evidência, é importante salientar que o estado da Califórnia é o que possui mais cursos de formação (n=14 - 15%), seguido por Nova Iorque com (n=13 - 14%) e Massachusetts (n=10 - 11%). De modo geral estes resultados eram esperados, levando-se em conta a tradição destes estados como verdadeiros polos de referência em tecnologia da informação, bem como pela grande concentração de *startups* e empresas no setor de comunicação, mídias, saúde, financeiro, entre outros interessados na contratação de cientistas de dados.

Do total de cursos, 58 (54%) são ofertados exclusivamente na modalidade presencial (*campus*), 28 (33%) online e o restante oferece opções híbridas (semipresencial) e adaptadas às necessidades dos estudantes. Grande parte dos cursos 65 (70%) são cursos em nível de mestrado, 21 (23%) certificados de estudos avançados, que no Brasil seria o equivalente a uma especialização ou habilitação, e apenas sete em nível de doutorado.

Há uma grande variedade nas denominações dos cursos, sendo que em alguns casos estes são específicos para um setor de aplicação como, por exemplo, o *Master of Science in Healthcare Data Analytics* pela *Clarkson*

University de *New York*, dedicado à área da saúde, ou mesmo combinado a áreas correlatas, como o *Master of Advanced Study in Data Science and Engineering* da *University of California - San Diego*. As universidades *University of California* e *Pennsylvania State University* são as que lideram em números de cursos, com cinco e quatro respectivamente, considerando seus diferentes *campi*.

Nenhum programa apresentava a data de implantação de forma evidente em seus *websites*. Pelo fato desta data fornecer um interessante indicador quanto à evolução da oferta em cursos para formação de cientistas de dados, enviamos um email solicitando essa informação pontual aos programas. Dos 93 emails totais, recebemos 62 (68%) respostas, sendo que duas respostas (1958 e 1961) foram descartadas para fins de análise, por obviamente se tratarem da implantação de cursos já extintos, substituídos ou que sofreram significativa transformação curricular e de nomenclatura, já que antecedem em décadas as questões relativas à ciência de dados. Uma terceira resposta de implantação foi excluída, pois era referente a um curso de Mestrado em Tecnologia da Informação com 12 diferentes áreas de concentração, podendo não refletir, portanto, a área de ciência de dados especificamente. Com base na amostra estudada e nas 60 (64,5%) respostas computadas, destacamos o curso *Master of Professional Studies (MPS) in Applied Statistics* com ênfase em *Data Science* pela *Cornell University* iniciado em 2008. A partir de 2011 surgiram os cursos *Master in Computational Data Science* pela *Chapman University* e o *Master in Business in Analytics and Data Science* pela *Rutgers University*, e houve um crescimento constante a partir de então. Quatro novos cursos foram implementados em 2012, nove em 2013, 18 em 2014, 14 em 2015 e 12 neste ano de 2016. Nota-se, portanto, com base nas informações prestadas por email, que o ápice de implantação de cursos em ciência de dados nos EUA foi em 2014.

Para conclusão dos programas analisados são necessários, em média, 32 créditos. Cabe ressaltar, no entanto, que os dados de créditos não foram localizados nos sites de quatro programas, e, portanto, os resultados são referentes a apenas 89 casos. Também cumpre frisar que há grande variedade

nos requisitos de créditos mínimos, que pode ser de apenas nove créditos, como nos casos dos do *Certificate in Data Science* oferecido na modalidade online pela *Washington University - Seattle*, ou chegar a até 144 créditos, como no caso do programa *Master of Computational Data Science* da *Carnegie Mellon University* na *Pennsylvania*.

Não foi possível identificar tempo de duração ou carga horária total para a grande maioria dos programas analisados. Apenas cerca de 5% das instituições provêm esta informação em seus *websites*. Isto pode ser explicado pelo fato dos cursos analisados serem em nível de pós-graduação, funcionarem por sistema de créditos e terem uma maior flexibilidade para a conclusão de seus programas.

Vistos os aspectos gerais de distribuição geográfica, tipologia de curso, tempo de existência e quantidade média de créditos exigidos para a conclusão dos programas, a seguir serão descritos quais atributos os programas tendem a solicitar dos ingressantes.

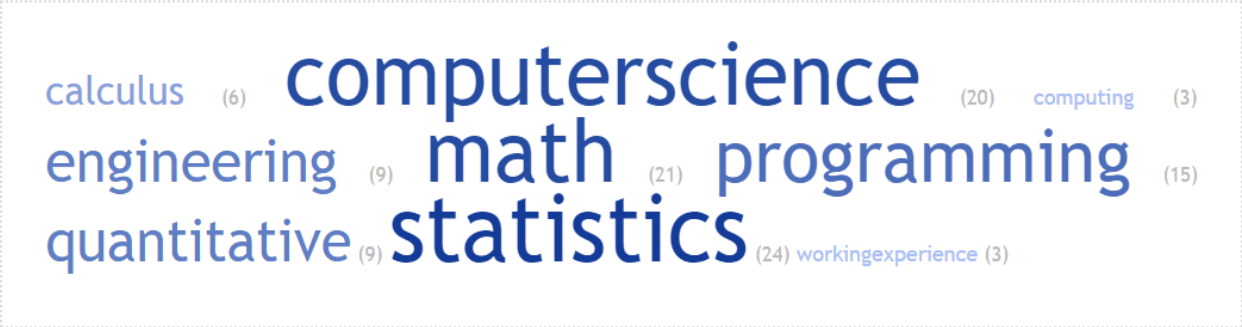
3.2 Dos Pré-requisitos para Ingresso

O sistema de ingresso em cursos de pós-graduação nos EUA, em seus diferentes níveis, exige dos candidatos uma série de documentações e testes que podem variar de instituição para instituição. Porém, de modo geral os candidatos devem apresentar os resultados de testes parametrizados como o *Graduate Management Admission Test* (GMAT), *Graduate Record Examination* (GRE) que seriam equivalentes a um vestibular unificado que testa os conhecimentos dos candidatos em nível superior, históricos escolares, cartas de intenção, currículo, cartas de recomendação, e realizar entrevistas e testes de proficiência em língua inglesa para os não-nativos.

Esta pesquisa teve como foco de análise a identificação dos diferentes pré-requisitos em termos de conhecimento prévio e de experiência profissional, que são descritos como exigidos e/ou desejáveis nos *websites* de instituições que ofertam cursos de formação em ciência de dados. Desse modo, parte-se do princípio de que com base nessas especificações é possível traçar um perfil geral e mínimo dos ingressantes nos programas avaliados.

Dentre os 93 programas, 56 (60%) exigem algum quesito específico de conhecimento prévio ou experiência aos ingressantes. A partir da padronização e refinamento dos termos-chave, foi possível identificar que a formação em nível de Bacharelado e/ou cursos e créditos cumpridos nas áreas de Estatística (43%), Matemática (37,5%), e na Ciência da Computação (36%) estão entre as mais requisitadas, conforme ilustrado na Figura 2:

Figura 2 - Pré-requisitos para ingresso nos cursos.



Fonte: Elaborado a partir de dados da pesquisa[7].

Em muitas das ocorrências, os programas requerem formação em Ciência da Computação, Estatística ou Matemática, em outros casos menos recorrentes os programas indicam formação na área de exatas de modo geral, incluindo as diferentes Engenharias. Como assinalado anteriormente no referencial teórico, as origens da ciência de dados estiveram balizadas nas ciências exatas, e estreitamente relacionada à estatística aplicada, o que explica, em parte, a existência destes pré-requisitos.

Habilidades de programação, cálculo, dados quantitativos em geral e diferentes níveis e tipos de análises estatísticas também são frequentemente requisitados pelos programas investigados, novamente reforçando a influência estatística, matemática e computacional de alguns cursos que exigem um conhecimento de partida mínimo como nivelamento, para que os estudantes consigam acompanhar as disciplinas e demais atividades do programa.

[7] Inclui somente pré-requisitos com pelo menos três ocorrências.

Observa-se, no entanto, que questões gerenciais e de negócios assinaladas como essenciais para o recrutamento de cientistas de dados não são estipuladas como pré-requisitos para o ingresso nos cursos. Todavia, em três casos, o quesito de experiência profissional constitui como critério para o ingresso nos cursos, como no caso do *Master of Science in Statistics - Data Science* pela *University of Wisconsin* em *Madison*, que solicita de três a cinco anos de experiência profissional, o *Master of Advanced Study in Data Science and Engineering* da U*niversity of California* em *San Diego*, que solicita dois anos de atuação no mercado de trabalho, e o *Online Hybrid Master of Sciences in Business Data Analytics* da *West Virginia University.*

A seção seguinte apresentará como os programas em ciência de dados delineiam suas propostas e metas ao seu público-alvo e à comunidade interessada.

3.3 Dos Objetivos e Competências Formativas Gerais

Conforme enunciado anteriormente, interessava-nos também entender como as instituições identificadas definem os objetivos de seus programas aos ingressantes potenciais e à comunidade em geral. Isso porque é por meio deste esclarecimento que os programas deliberadamente contemplam sua missão e o perfil de profissional que pretendem formar, delineando as competências formativas gerais do curso.

A análise textual dos objetivos dos programas permitiu classificá-los em três níveis. Há aqueles que são mais genéricos e apenas indicam a necessidade de preparar profissionais para o competitivo mercado de trabalho, ou tão somente para a formação e obtenção do título que o programa oferece, o que ocorreu em 25 (27%) casos. Existem aqueles que buscam contextualizar as demandas da era *big data* e as oportunidades para atuação neste cenário indicando a crescente demanda do mercado de trabalho, o que ficou evidente nos objetivos de seis (6%) dos programas analisados. Por fim, os demais 62 (67%) programas seguem uma abordagem mais detalhada de seus objetivos indicando diferentes domínios e *expertises* esperados dos egressos, muitos dos quais voltados para os aspectos computacionais, estatísticos e de domínio

ou orientado ao aspecto gerencial e de negócios, conforme indicado pela literatura. Nota-se forte ênfase desses objetivos, principalmente em torno dos seguintes eixos temáticos:

- Análise estatística avançada (preditiva e inferencial);
- Aprendizagem de máquina (*machine learning*);
- Capacidade gerencial e de tomada de decisão subsidiada por dados;
- Computação aplicada e programação (*applied computing and programming*);
- Econometria;
- Gerenciamento e curadoria de dados;
- Identificação de padrões e *insights* por mineração de dados (*data mining*);
- Inteligência de Negócios (*business intelligence*);
- Modelagem de dados;
- Processamento de linguagem natural (*natural language processing*);
- Segurança de dados, cibersegurança e privacidade de dados (*data privacy*);
- Uso e desenvolvimento de ferramentas analíticas;
- Visualização e representação gráfica de dados.

Vale ressaltar que grande parte dos programas reforça o caráter de aplicação dos conteúdos para a solução de problemas reais no contexto dos negócios e das organizações, tanto da iniciativa pública como privada. Isso fica explícito em pelo menos 13 programas que utilizam a expressões como "*real-world problems*", "*realistic settings*", "*realistic circumstances*" ou "*real-life problems*".

A análise ainda permitiu identificar que três programas destacam ênfase em aspectos éticos e de uso responsável de dados, o *Master in Data Science* do *Illinois Institute of Technology no estado de Illinois*, o *Graduate Certificate in Data Science* da *Regis University* no Colorado e o *Master of Science in Data Analytics* da *Southern New Hampshire University* em estado de mesmo nome. Essas habilidades são essenciais para aqueles que lidam com dados sensíveis e que devem preservar a identidade e a privacidade dos envolvidos.

4 IMPLICAÇÕES E CONSIDERAÇÕES FINAIS

Os resultados da pesquisa oferecem indicadores iniciais e um panorama geral sobre a composição dos cursos em ciência de dados dos EUA, dando subsídios para uma reflexão quanto ao público-alvo destes programas, seus objetivos, e diretrizes gerais a respeito da formação de cientistas de dados.

Embora ainda seja carreira recente em território nacional, e os anúncios ativos pelo *LinkedIn* no Brasil não ultrapassem uma centena, esta área vem crescendo no país. Mesmo que ainda de forma tímida, conforme destaca Breternitz, Lopes e Silva (2015), já existem algumas iniciativas no Brasil na modalidade *lato sensu* de cursos voltados para a ciência de dados oferecidos pela Universidade Presbiteriana Mackenzie, a Escola Superior de Propaganda e Marketing e a Fundação Getúlio Vargas. Em outros casos, como o curso de graduação em Ciência da Informação da Universidade Federal de Santa Catarina (UFSC), também observamos uma tendência em preparar profissionais com habilidades e competências semelhantes às comuns aos cientistas de dados. Desse modo, esperamos que os dados aqui relatados possam contribuir para a formação em ciência de dados em âmbito nacional, não necessariamente nos moldes do sistema estadunidense, mas considerando experiências pioneiras e tendências precursoras.

Como visto, tem havido uma expansão dos cursos destinados a formarem e a capacitarem a nova geração de cientistas de dados principalmente em nível de mestrado. Mais da metade dos cursos analisados são ofertados na modalidade presencial, embora haja uma parcela de cursos *online* ou híbridos. Há variações de nomenclatura e de créditos exigidos para conclusão destes cursos, porém há certa convergência de foco de interesse em estudantes com bagagem nas áreas da computação, a estatística e a matemática, indicando uma predominância de perfil do público-alvo e dos ingressantes destes programas. Os pré-requisitos solicitados aos ingressantes têm consonância com os objetivos e propostas gerais de abordagem de conteúdos dos cursos, que estão fortemente ligados às áreas acima indicadas. Também foi possível observar que as habilidades e competências exigidas em

vagas de empregos para esse profissional estão de certo modo contempladas nos objetivos e propostas de formação dos cursos.

A análise preliminar empreendida neste artigo oferece base para os percursos futuros da pesquisa, que buscará aprofundar a discussão acerca da formação curricular dos cientistas de dados, a partir da avaliação de documentos complementares (ementas, programas de disciplinas, projetos pedagógicos, entre outros). Tais documentos serão solicitados diretamente aos programas, haja vista que, em muitos casos, as informações prestadas nos *websites* são incipientes, ou restritas à comunidade interna.

REFERÊNCIAS

BIEHN, N. The missingV's in big data: viabilityandvalue. Wired, New York, 2013. Disponível em: <https://www.wired.com/insights/2013/05/the-missing-vs-in-big-data-viability-and-value>. Acesso em: 14 ago. 2016.

CHATFIELD, A. T. *et al*. Data Scientists as a game changers in big data environments. In: PROCEEDINGS OF THE 25TH AUSTRALASIAN CONFERENCE ON INFORMATION SYSTEMS (ACIS), *Anais…*, Auckland: Auckland Universityof Technology, 2014. p.1-11.

CLEVELAND, W. S. Data Science: anactionplan for expandingthetechnicalareasofthefieldofstatistics. *InternationalStatisticalReview*, Malden, MA, v. 69, p. 21-26, 2001. doi:10.1111/j.1751-5823.2001.tb00477.x

CONAWAY, D. *The data sciencevenndiagram*. 2010. Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram>. Acesso em: 10 ago. 2016.

CROWDFLOWER. *Data Science Report.* 2016. Disponível em: <http://visit.crowdflower.com/rs/416-ZBE-142/images/CrowdFlower_DataScienceReport_2016.pdf>. Acesso em: 10 set. 2016.

FINZER, W. The data scienceeducationdilemma. *Technology Innovations in StatisticsEducation*, Caifórnia, v. 7, n. 2, 2013. Disponível em: <http://escholarship.org/uc/item/7gv0q9dc>. Acesso em: 22 ago. 2016.

GRANVILLE, V. *Developinganalytictalent*: becoming a data scientist. Indianapolis: John Wiley, 2014.

KIM, J. Y.; LEE, C. K. Anempiricalanalysisofrequirements for data scientistsusing online jobpostings. *InternationalJournalof Software Engineeringand its Applications*, Seoul, v. 10, n. 4, p.161-172, 2016.

LANEY, D. 3D Data management: controlling data volume, velocityandvariety. *Application Delivery Strategies*, Stanford. 2001. Disponível em: <http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>. Acesso em: 20 ago. 2016.

LOUKIDES, Mike. *Whatis data science?*Sebastopol, CA: O'Reilly Media, 2011.

PATIL, T. H.; DAVENPORT, D. J. Data Scientist: thesexiestjobofthe 21st century. *Harvard Business Review*, Brighton, MA, 2012. Disponível em: <https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century>. Acesso em: 5 ago. 2016.

STANTON, J. *et al*. Interdisciplinary data scienceeducation. In: XIAO, N..; MCEWEN,L. R. *SpecialIssues in Data Management*. Washington, DC: American ChemicalSociety, 2012. p. 97-113. (ACS Symposium Series, v. 1110).doi: 10.1021/bk-2012-1110.ch006

SWAN, A.; BROWN, S. *The skills, role andcareerstructureof data scientistsandcurators*: anassessmentofcurrentpracticeand future needs. Reporttothe Joint Information Systems Committee (JISC). Truro: Key Perspectives for JISC, 2008. 34 p.

WARD, S.; BARKER, A. Undefinedby data: a surveyof big data definitions. 2013. Disponível em: <arXivpreprint arXiv:1309.5821>. Acesso em: 10 out. 2016.

VAN DER AALST, W. M. P. Data scientist: theengineerofthe future. In: *Enterprise Interoperability VI:* interoperability for agility, resilienceandplasticityofcollaborations. Springer: New York,2014. doi: 10.1007/978-3-319-04948-9_2

Title

Data science education**:** a preliminary analysis of the U.S landscape

Abstract
Introduction: Data scientists has received great attention in recent years following the demands of the labor market stimulated by the open science and big data era. Originally widespread in 2008 and, since then, present in many different industries and applications; data science was announced in 2012 as the most attractive and one of the best paid jobs of the century, culminating with an increasing supply of training courses.

Objective: Characterize and understand the formative aspects of data scientists.
Methodology: This article describes part of a survey research based on analysis of 93 degrees in data science offered by US institutions.
Results: The content analysis of the information publicized on the websites of the identified programs provides evidence that this professional is trained to deal with issues related to the collection, treatment, processing, analysis, visualization and curation of large and heterogeneous data collections in order to solving real-life and practical problems.
Conclusion: Findings also revealed that, in general, training in science data places great emphasis on statistical skills, mathematics and computing, including programming and advanced modeling, many of which are placed as prerequisites for admission in these programs.

Keywords: Data Science. Data Scientist. Professional Skills. Professional Qualification.

Titulo

Formación en ciencias de datos: un análisis preliminar de las perspectivas de EE.UU.

Resumen
Introducción: Los Científicos de Datos han recibido una gran importancia en los últimos años a raíz de las demandas del mercado de trabajo estimuladas por la ciencia abierta y la era de grandes volúmenes de datos. Ampliamente publicada en 2008, y ahora presente en diferentes sectores y aplicaciones, la terminología "científico de datos" se anunció en 2012 como la más atractiva y uno de los mejor pagados del siglo XXI, que culminó con una creciente oferta de cursos de formación.
Objetivo: Caracterizar y entender los aspectos formativos del científico de datos.
Metodología: En el artículo se relata el recorte de un estudio de investigación basado en un análisis preliminar de 93 cursos en ciencia de datos que ofrecen las instituciones de los Estados Unidos.
Resultados: El análisis del contenido de la información contenida en los sitios web de los programas identificados ha puesto de manifiesto que este profesional está capacitado para hacer frente a cuestiones relacionadas con la recolección, tratamiento, procesamiento, análisis, visualización y la curaduría de grandes y heterogéneas colecciones de datos orientadas a la resolución de problemas prácticos y reales.
Conclusión: Se constató que, en general, la formación en la ciencia de datos concede gran énfasis en las habilidades de estadísticas, matemáticas e computacionales, incluyendo la programación y modelado avanzado, muchos de los cuales son requisitos previos para la admisión a estos cursos.

Palabras clave: Ciencia de Datos. Cientista de Datos. Competencias profesionales. Formación professional.

Enviado em: 17.07.2016.
Aceito em: 20.11.2016.

APÊNDICE A

Cursos Analisados

Nome do Curso	Instituição	Estado
Master of Science in Data Science	Galvanize University – New Haven	California
Online Masters of Information and Data Science	University of California - Berkeley	California
Master of Engineering - Concentration in Data Science & Systems	University of California - Berkeley	California
Certificate in Data Science	California State University - Fullerton	California
Doctorate in Computational and Data Science	Chapman University	California
Master of Computational and Data Science	Chapman University	California
Master of Business Administration with concentration in Data Science and Business Analytics	Santa Clara University	California
Masters of Science in Information Systems & Technology: Concentration in Data Science & Analytics	Claremont Graduate University	California
Master of Science in Statistics: Data Science	Stanford University	California
Data Science Certificate	University of California - Irvine	California
Master of Science in Engineering - Data Science Specialization	University of California - Riverside	California
Master of Advanced Study in Data Science and Engineering	University of California - San Diego	California
Master of Science in Computer Science - Data Science	University of Southern California	California
Doctor of Philosophy in Data Sciences & Operations	University of Southern California	California
Graduate Certificate in Data Science	Regis University	Colorado
Master in Data Science	New College of Florida	Florida
Doctor of Philosophy in Analytics and Data Science	Kennesaw State University	Georgia
Online Master of Science in Data Science	Lewis University	Illinois
Graduate Certificate in Data Science	Elmhusrt College	Illinois
Master of Science in Data Science	Elmhusrt College	Illinois
Master of Data Science	Illinois Institute of Technology	Illinois
Master of Computer Science in Data Science	University of Illinois at Urbana-Champaign	Illinois
Online Master of Science in Data Science	Saint Mary's College	Indiana
Master of Science in Data Science	Indiana University Bloomington	Indiana
Online Certificate in Data Science	Indiana University Bloomington	Indiana
Certificate in Applied Econometric and Data Science Foundations using SAS	Valparaiso University	Indiana
Master of Science in Applied Data Science	Bay Path University	Massachusetts
Data Science Certificate	Harvard University	Massachusetts
Master of Science in Computer Science with Concentration in Data Science	University of Massachusetts	Massachusetts
Doctor of Philosophy in Business Administration - Information Systems for Data Science Track	University of Massachusetts - Boston	Massachusetts
Graduate Certificate in Data Science	Worcester Polytechnic Institute	Massachusetts
Master of Science in Data Science	Worcester Polytechnic Institute	Massachusetts
Doctor of Philosophy in Data Science	Worcester Polytechnic Institute	Massachusetts
Graduate Data Science Certificate Program	University of Michigan - Ann Arbor	Michigan
Master's of Science in Data Science	University of Minnesota - Twin Cities	Minnesota
Master of Science in Data Science	University of St. Thomas	Minnesota
Graduate Certificate in Business Analytics and Data Science	Missouri University of Science and Technology	Missouri
Data Science and Business Analytics Certificate	Rockhurst University	Missouri
Data Science and Business Intelligence Certificate	Rockhurst University	Missouri
Master of Business and Science degree in Analytics - discovery informatics & data sciences	Rutgers University	New Jersey
Master of Science in Data Science with a concentration in Business Analytics	Saint Peter's University	New Jersey
Certificate of Advanced Study in Data Science	Syracuse University	New York
Master of Science in Data Science	Columbia University in the City of New York	New York
Master of Professional Studies in Applied Statistics (Option II: Data Science)	Cornell University	New York

Master of Science in Data Science	New York University	New York
Master of Science in Information Technology - Concentration in Data Science and Analytics	Rensselaer Polytechnic Institute	New York
Master of Science in Data Science	University of Rochester	New York
Graduate Certificate in Data Science and Business Analytics	University of North Carolina at Charlotte	North Carolina
Professional Science Master's in Data Science and Business Analytics	University of North Carolina at Charlotte	North Carolina
Master of Data Science and Analytics	University of Oklahoma Norman Campus	Oklahoma
Master of Computational Data Science	Carnegie Mellon University	Pennsylvania
Master of Science in Data Science	Mercyhurst University	Pennsylvania
Master of Science in Data Science	South Dakota State University	South Dakota
Online Master of Science in Data Science	Southern Methodist University	Texas
Master of Science in Data Science	Texas Tech University	Texas
Master of Science in Data Science	University of Virginia	Virginia
Certificate in Data Science	University of Washington - Seattle Campus	Washington
Doctor of Philosophy in Big Data and Data Science	University of Washington - Seattle Campus	Washington
Master of Science in Data Science	George Washington University	Washington, D.C.
Certificate in Data Science	Georgetown University	Washington, D.C.
Master of Science in Analytics, Concentration in Data Sciences	Georgetown University	Washington, D.C.
Online Master of Science in Data Science	University of Wisconsin Colleges	Wisconsin
Master of Science in Statistics - Data Science	University of Wisconsin - Madison	Wisconsin
Master of Science in Data Analytics	National University	California
Doctor of Computer Science - Concentration in Big Data Analytics	Colorado Technical University	Colorado
Master of Science in Data Analytics	University of Central Florida	Florida
Master of Science – Certificate of Specialization in Data Analytics	Illinois Institute of Technology	Illinois
Master of Science in Data Analytics	University of Maryland - University College	Maryland
Data Analytics Graduate Certificate	Boston University	Massachusetts
Master of Science in Computer Information Systems - Data Analytics Concentration	Boston University	Massachusetts
Graduate Certificate in Data Analytics	Northeastern University	Massachusetts
Online Accelerated Master of Business Administration - Data Analytics	Saint Mary's University of Minnesota	Minnesota
Master of Science in Information Systems - Data Analytics Track	University of Nevada - Reno	Nevada
Master of Science in Data Analytics	Southern New Hampshire University	New Hampshire
Master of Business Administration in Data Analytics	Thomas Edison State University	New Jersey
Master of Science in Healthcare Data Analytics	Clarkson University	New York
Online Master of Science in Data Analytics	CUNY Graduate School and University Center	New York
Master of Arts in Data Analytics & Applied Social Research	CUNY Queens College	New York
Master of Science in Data Analytics	Fordham University	New York
Master of Science in Applied Mathematics - Data Analytics	Manhattan College	New York
Advanced Certificate in Big Data Analytics	Rochester Institute of Technology	New York
Professional Master of Science in Computer Science (Concentration in Data Analytics)	University of Rochester	New York
Master of Information Systems Management, Business Intelligence and Data Analytics	Carnegie Mellon University	Pennsylvania
Master of Science in Information Technology, Business Intelligence and Data Analytics	Carnegie Mellon University	Pennsylvania
Master of Professional Studies in Data Analytics - Business Analytics Option	Pennsylvania State University - Main Campus	Pennsylvania
Graduate Certificate in Data Analytics	Pennsylvania State University - Penn State Great Valley	Pennsylvania
Master of Professional Studies in Data Analytics	Pennsylvania State University - Penn State Great Valley	Pennsylvania
Master of Professional Studies in Data Analytics	Pennsylvania State University - World Campus	Pennsylvania

Master of Science in Information Science - Big Data Analytics	University of Pittsburgh - Pittsburgh Campus	Pennsylvania
Master of Science in Applied Statistics and Data Analytics	Southern Methodist University	Texas
Master of Science in Data Analytics	The University of Texas at San Antonio	Texas
Master of Science in Data Analytics Engineering	George Mason University	Virginia
Online Hybrid MS in Business Data Analytics	West Virginia University	West Virginia