**Syllabus**

**Course Syllabus**

**Defining Data Science and What Data Scientists Do**

* Defining Data Science
* What is Data Science?
* Fundamentals of Data Science
* The Many Paths to Data Science
* Advice for New Data Scientists
* Data Science: The Sexiest Job in the 21st Century

**What Do Data Scientists Do?**

* A day in the Life of a Data Scientist
* Old problems, new problems, Data Science solutions
* Data Science Topics and Algorithms
* What is the cloud?
* What Makes Someone a Data Scientist?

**Data Science Topics**

* Foundations of Big Data
* How Big Data is Driving Digital Transformation
* What is Hadoop?
* Data Science Skills & Big Data
* Data Scientists at New York University
* Data Mining
* Quiz: Data Mining

**Deep Learning and Machine Learning**

* What's the difference?
* Neural Networks and Deep Learning
* Applications of Machine Learning
* Regression
* Quiz: Regression

**Data Science in Business**

* Applications of Data Science
* How Data Science is Saving Lives
* How Should Companies Get Started in Data Science?
* Applications of Data Science
* The Final Deliverable
* Quiz: The Final Deliverable

**Careers and Recruiting in Data Science**

* How Can Someone Become a Data Scientist?
* Recruiting for Data Science
* Careers in Data Science
* High School Students and Data Science Careers

**The Report Structure**

* The Report Structure
* Quiz: The Report Structure
* Final Assignment

# Professional Certificate Career Support

## Coursera Community and Career Support

As a Data Science learner on Coursera, you have access to networking opportunities in Coursera’s [Professional Certificate Community](http://bit.ly/2Ij0Kg9) and [Data Science](http://bit.ly/2VEY9VI) forums. Talk about what you’re learning, ask questions, find peers to work with on projects, and share your career goals.

## Post-Completion Career Support Services for Professional Certificates

Completing a Professional Certificate on Coursera unlocks access to a private Professional Certificate Alumni Resources community, which provides exclusive career support resources, including:

* Step-by-step **guide to ensure your success** at every stage of your job search.
* 1 year of **free access** to Big Interview’s expert video lessons, resume builder, and interactive interview practice tools (a $79/month value).
* **A network** and support of fellow completers of Coursera Professional Certificates.
* A variety of **special offers** such as career coaching, webinars, and more.

After completing your Professional Certificate, you’ll get an email telling you how to access these career support resources.

Questions? You can also always contact Coursera Careers Team by emailing [career-support@coursera.org](mailto:professional-certificate-career-support@coursera.org).

## Defining Data Science

# What is Data Science?

Ciência de dados é um processo, não um evento.

É o processo de utilização de dados para entender coisas diferentes,

para entender o mundo.

Para mim é quando você tem um modelo ou hipótese de um problema,

e você tenta validar essa hipótese ou modelo com base em seus dados.

Ciência de dados é a arte de descobrir

as ideias e tendências que se escondem atrás dos dados.

É quando você traduz dados em uma história.

Então usa narrativas para gerar percepções.

E com essas percepções,

Você pode fazer escolhas estratégicas para uma empresa ou instituição.

Ciência de dados é um campo sobre processos e sistemas para

extrair dados de várias formas, estruturados ou não.

Ciência de dados é o estudo dos dados.

Como ciências biológicas é um estudo de biologia,

Ciência física, é o estudo de reações físicas.

Dados são reais, dados tem propriedades,

e precisamos estudá-los, se vamos trabalhar neles.

Ciência de dados envolve dados e alguns sinais.

A definição ou o nome surgiu

nos anos 80 e 90, quando alguns professores analisavam o currículo de estatísticas,

e pensaram que seria melhor chamá-lo de ciência de dados.

Mas o que é ciência de dados?

Eu veria ciência de dados como uma tentativa de trabalhar com dados,

para encontrar respostas a perguntas que estão sendo exploradas.

Em poucas palavras, é mais sobre dados do que sobre ciência.

Se você tiver os dados, e se você tiver curiosidade,

e se você estiver trabalhando com os dados,

e manipulando-os, explorando-os,

esse exercício de analisar os dados,

na tentativa de obter algumas respostas é ciência de dados.

Ciência de dados é relevante hoje porque temos toneladas de dados disponíveis.

Costumávamos nos preocupar com falta de dados.

Agora temos um dilúvio de dados.

No passado, não tínhamos algoritmos, agora temos algoritmos próprios.

No passado, os programas eram caros,

Agora eles são abertos e gratuitos.

No passado, nós não podíamos armazenar grandes quantidades de dados,

Agora, por uma fração do custo,

Nós podemos ter zilhões de conjuntos de dados para um custo muito baixo.

Assim, as ferramentas para trabalhar com dados,

a variabilidade dos dados,

e a capacidade de armazenar e analisar dados,

está muito em conta, está muito disponível,

é universal, é o presente.

Nunca houve um melhor momento para ser um cientista de dados.

# Fundamentals of Data Science

Todos os que você pedir lhe darão uma descrição ligeiramente diferente do que

é a Ciência de Dados, mas a maioria das pessoas concorda que ela tem um componente significativo de análise de dados. A análise de dados não é

nova. O que é novo é a grande quantidade de dados disponíveis a partir de fontes extremamente variadas: de

arquivos de log, e-mail, mídia social, dados de vendas, arquivos de informações de pacientes,

dados de desempenho esportivo, dados de sensores, câmeras de segurança e muito mais. Ao mesmo tempo que

há mais dados disponíveis do que nunca, temos o poder computacional necessário para fazer uma

análise útil e revelar novos conhecimentos. A ciência de dados pode ajudar as organizações a entender

seus ambientes, analisar problemas existentes e revelar oportunidades anteriormente ocultas.

Os cientistas de dados usam a análise de dados para adicionar ao conhecimento da organização, investigando

os dados, explorando a melhor maneira de usá-los para fornecer valor aos negócios.

Então, qual é o processo da ciência de dados? Muitas organizações usarão a ciência de dados para se

concentrar em um problema específico, e por isso é essencial esclarecer a pergunta que a organização

quer que seja respondida. Este primeiro e mais crucial passo define como o projeto de ciência de dados

avança. Bons cientistas de dados são pessoas curiosas que fazem perguntas para esclarecer a

necessidade do negócio. As próximas perguntas são: “Que dados precisamos

para resolver o problema e de onde esses dados virão?”. Os cientistas de dados podem analisar

dados estruturados e não estruturados de muitas fontes e, dependendo da natureza do

problema, podem optar por analisar os dados de diferentes maneiras. O uso de vários modelos para

explorar os dados revela padrões e outliers; às vezes, isso confirmará o que a organização

suspeita, mas às vezes será um conhecimento completamente novo, levando a organização a

uma nova abordagem. Quando os dados revelam suas percepções, o

papel do cientista de dados torna-se o de um contador de histórias, comunicando os resultados às

partes interessadas do projeto. Os cientistas de dados podem usar ferramentas poderosas de visualização de dados

para ajudar as partes interessadas a entender a natureza dos resultados e as ações recomendadas

a serem tomadas. A Ciência de Dados está mudando a maneira como trabalhamos;

está mudando a maneira como usamos os dados e está mudando a maneira como as organizações entendem

o mundo.

# The Many Paths to Data Science

[SOM]

[MÚSICA]

Ciência de dados realmente não existia quando eu estava crescendo.

Não é algo que eu já acordei e

disse: “Quero ser cientista de dados quando crescer.

Não, não existia.

Eu não sabia que estaria trabalhando em ciência de dados.

Quando eu cresci, não existe esse campo chamado ciência de dados.

E eu acho que é realmente novo.

A ciência de dados não existia até 2009, 2011.

Alguém como DJ Patil ou Andrew Gelman inventou o termo.

Antes disso, havia estatísticas.

E eu não queria ser nenhum desses.

Eu queria estar no negócio.

E então achei a ciência de dados muito mais interessante.

Estudei estatísticas, foi assim que comecei.

Reproduza o vídeo começando em ::45 e siga a transcrição0:45

Passei por muitas fases diferentes da minha vida onde queria ser cantora e

depois médica.

E então eu percebi que eu era bom em matemática.

Então eu escolhi uma área que estava focada em análise quantitativa.

E a partir daí eu acho que eu queria trabalhar com dados.

Não necessariamente ciência de dados, como é conhecida hoje.

A primeira vez que tive contato com ciência de dados,

quando eu era meu primeiro ano como engenharia mecânica.

E as empresas de consultoria estratégica usam a ciência de dados para tomar decisões.

Então, foi meu primeiro contato com a ciência de dados.

Eu tinha um problema complicado que eu precisava resolver, e

as técnicas usuais que tínhamos na época não podiam ajudar com esse problema.

Eu me formei com um diploma de matemática no pior momento possível,

logo após a crise econômica, e você tinha que ser útil para conseguir um emprego.

Então eu fui e me formei em estatística.

E então eu trabalhei o suficiente que eram chamados de cientista de dados que de

repente eu me tornei um.

Minha graduação era em negócios, e me formei em política

, filosofia e economia.

E depois fiz um mestrado em análise de negócios na

Universidade de Nova Iorque na Stern School of Business.

Quando saí da minha graduação, a primeira empresa em que entrei, descobriu-se

que eles estavam analisando dados eletrônicos de pontos de venda para fabricantes de varejo.

E o que estávamos fazendo era ciência de dados.

Mas só começamos a usar esse termo muito mais tarde.

Na verdade, eu diria que há quatro ou cinco anos foi quando começamos a chamá-lo de análise e

ciência de dados.

Eu tinha várias opções para o meu estágio aqui no Canadá.

E uma das opções era trabalhar com ciência de dados.

Costumava trabalhar com desenvolvimento de projetos.

Mas acho que foi uma boa escolha.

E então eu começo meu estágio com ciência de dados.

Sou engenheiro civil por treinamento, então todos os engenheiros trabalham com dados.

Eu diria que o uso convencional da

ciência de dados em minha vida começou com a pesquisa de transporte.

Comecei a construir grandes modelos tentando prever o tráfego nas ruas,

tentando determinar o congestionamento e emissões de gases com efeito de estufa ou emissões de escape.

Então eu acho que foi onde meu começo foi.

E comecei a construir estes modelos quando era estudante de pós-graduação

na Universidade de Toronto.

Começou a trabalhar com conjuntos de dados muito grandes, olhando para amostras domésticas de,

digamos, 150.000 domicílios de meio milhão de viagens.

E isso, também, estou falando de meados dos anos 90 quando este

deveria ser um conjunto de dados muito grande, mas não nos termos de hoje.

Mas foi assim que comecei.

Continuei a trabalhar com ele.

Depois mudei-me para a Universidade McGill, onde era professor de

engenharia de transportes.

E criei modelos de dados ainda maiores que envolviam dados e análises.

E então eu diria, sim, pesquisa de transporte me trouxe para ciência de dados.

[ MUSIC]

[SOM]

[MÚSICA]

Ciência de dados realmente não existia quando eu estava crescendo.

Não é algo que eu já acordei e

disse: “Quero ser cientista de dados quando crescer.

Não, não existia.

Eu não sabia que estaria trabalhando em ciência de dados.

Quando eu cresci, não existe esse campo chamado ciência de dados.

E eu acho que é realmente novo.

A ciência de dados não existia até 2009, 2011.

Alguém como DJ Patil ou Andrew Gelman inventou o termo.

Antes disso, havia estatísticas.

E eu não queria ser nenhum desses.

Eu queria estar no negócio.

E então achei a ciência de dados muito mais interessante.

Estudei estatísticas, foi assim que comecei.

Reproduza o vídeo começando em ::45 e siga a transcrição0:45

Passei por muitas fases diferentes da minha vida onde queria ser cantora e

depois médica.

E então eu percebi que eu era bom em matemática.

Então eu escolhi uma área que estava focada em análise quantitativa.

E a partir daí eu acho que eu queria trabalhar com dados.

Não necessariamente ciência de dados, como é conhecida hoje.

A primeira vez que tive contato com ciência de dados,

quando eu era meu primeiro ano como engenharia mecânica.

E as empresas de consultoria estratégica usam a ciência de dados para tomar decisões.

Então, foi meu primeiro contato com a ciência de dados.

Eu tinha um problema complicado que eu precisava resolver, e

as técnicas usuais que tínhamos na época não podiam ajudar com esse problema.

Eu me formei com um diploma de matemática no pior momento possível,

logo após a crise econômica, e você tinha que ser útil para conseguir um emprego.

Então eu fui e me formei em estatística.

E então eu trabalhei o suficiente que eram chamados de cientista de dados que de

repente eu me tornei um.

Minha graduação era em negócios, e me formei em política

, filosofia e economia.

E depois fiz um mestrado em análise de negócios na

Universidade de Nova Iorque na Stern School of Business.

Quando saí da minha graduação, a primeira empresa em que entrei, descobriu-se

que eles estavam analisando dados eletrônicos de pontos de venda para fabricantes de varejo.

E o que estávamos fazendo era ciência de dados.

Mas só começamos a usar esse termo muito mais tarde.

Na verdade, eu diria que há quatro ou cinco anos foi quando começamos a chamá-lo de análise e

ciência de dados.

Eu tinha várias opções para o meu estágio aqui no Canadá.

E uma das opções era trabalhar com ciência de dados.

Costumava trabalhar com desenvolvimento de projetos.

Mas acho que foi uma boa escolha.

E então eu começo meu estágio com ciência de dados.

Sou engenheiro civil por treinamento, então todos os engenheiros trabalham com dados.

Eu diria que o uso convencional da

ciência de dados em minha vida começou com a pesquisa de transporte.

Comecei a construir grandes modelos tentando prever o tráfego nas ruas,

tentando determinar o congestionamento e emissões de gases com efeito de estufa ou emissões de escape.

Então eu acho que foi onde meu começo foi.

E comecei a construir estes modelos quando era estudante de pós-graduação

na Universidade de Toronto.

Começou a trabalhar com conjuntos de dados muito grandes, olhando para amostras domésticas de,

digamos, 150.000 domicílios de meio milhão de viagens.

E isso, também, estou falando de meados dos anos 90 quando este

deveria ser um conjunto de dados muito grande, mas não nos termos de hoje.

Mas foi assim que comecei.

Continuei a trabalhar com ele.

Depois mudei-me para a Universidade McGill, onde era professor de

engenharia de transportes.

E criei modelos de dados ainda maiores que envolviam dados e análises.

E então eu diria, sim, pesquisa de transporte me trouxe para ciência de dados.

[ MUSIC]

# Advice for New Data Scientists

Meu conselho a um aspirante a cientista de dados é ser curioso,

extremamente argumentativo e crítico.

Curiosidade é absoluta necessidade.

Se você não estiver curioso, você não saberia o que fazer com os dados.

Julgamento porque se você não tem

noções preconcebidas sobre as coisas você não saberia por onde começar.

Argumentativo porque se você pode argumentar e se você pode argumentar um caso,

pelo menos você pode começar em algum lugar e então você aprende com dados e, em seguida, você

modifica suas suposições e hipóteses e seus dados iria ajudá-lo a aprender.

E você pode começar no ponto errado.

Você pode dizer que eu pensei que eu acreditava nisso,

mas agora com dados eu sei disso.

Então, isso permite que você um processo de aprendizagem.

Então, curiosidade ser capaz de tomar uma posição,

posição forte, e, em seguida, avançar com ela.

A outra coisa que o cientista de dados precisaria é de

algum conforto e flexibilidade com plataformas de análise: algum software,

alguma plataforma de computação, mas isso é secundário.

O mais importante é a curiosidade e a capacidade de tomar posições.

Uma vez que você tenha feito isso, uma vez que você tenha analisado,

então você tem algumas respostas.

E essa é a última coisa que um cientista de dados precisa,

e essa é a capacidade de contar uma história.

Que uma vez que você tem suas análises,

uma vez que você tem suas tabulações,

agora você deve ser capaz de contar uma grande história a partir dele.

Porque se você não contar uma grande história a partir dela,

suas descobertas permanecerão escondidas,

permanecerão enterradas, ninguém saberia.

Mas a sua ascensão à proeminência está basicamente confiando na sua habilidade de contar grandes histórias.

Um ponto de partida seria ver qual é a sua vantagem competitiva.

Você quer ser um cientista de dados em qualquer campo ou em um campo específico?

Porque, digamos que você queira ser um cientista de dados e trabalhar para

uma empresa de TI ou uma empresa baseada na Web ou na Internet,

então você precisa de um conjunto diferente de habilidades.

E se você quiser ser um cientista de dados no setor da saúde,

então você precisa de diferentes conjuntos de habilidades.

Então descubra primeiro o que você está interessado,

e qual é a sua vantagem competitiva.

Sua vantagem competitiva não será necessariamente suas habilidades analíticas.

Sua vantagem competitiva é sua compreensão de algum aspecto da

vida onde você excede os outros em entender isso.

Talvez seja filme, talvez seja varejo,

talvez saúde, talvez seja computadores.

Uma vez que você descobriu onde está sua experiência,

então você começa a adquirir habilidades analíticas.

Quais plataformas aprender e essas plataformas,

essas ferramentas seriam específicas para o setor em que você está interessado.

E então, uma vez que você tenha alguma proficiência nas ferramentas,

a próxima coisa seria aplicar suas habilidades a problemas reais,

e então dizer ao resto do mundo o que você pode fazer com ele.

Leitura

## Parabéns! Você foi aprovado!

Nota recebida 86,66%

Nota do envio mais recente 86,67%

Para ser aprovado 66% ou superior

Ir para o próximo item

### 1.

Pergunta 1

Walmart addressed its analytical needs by approaching Kaggle to host a competition for analyzing its proprietary data.

1 / 1 ponto



False.



True.

Correto

Correct.

### 2.

Pergunta 2

What is the average base salary of a data scientist reported by the New York Times?

1 / 1 ponto



$100,000



$150,000



$112,000



$16 per hour



$85,000 + Bonus

Correto

### 3.

Pergunta 3

According to professor Haider, the three important qualities to possess in order to succeed as a data scientist are:

0.6 / 1 ponto



Good at Math and Statistics.



Judgemental.

Correto

According to professor Haider, the three important qualities are being curious, judgemental, and argumentative.



Good Story Teller (Argumentative).



Proficient in Programming.

Não deve ser selecionado

According to professor Haider, the three important qualities are being curious, judgemental, and argumentative, and although being proficient in programming is a very important skill, but it is not one of important qualities that professor Haider lists.



Curious.

Correto

According to professor Haider, the three important qualities are being curious, judgemental, and argumentative.

# Lesson Summary

In this lesson, you have learned:

* Data science is the study of large quantities of data, which can reveal insights that help organizations make strategic choices.
* There are  many paths to a career in data science; most, but not all, involve a little math, a little science, and a lot of curiosity about data.
* New data scientists need to be curious, judgemental and argumentative.
* Why data science is considered the sexiest job in the 21st century, paying high salaries for skilled workers.

## What Do Data Scientists Do?

# A day in the Life of a Data Scientist

Eu construí um mecanismo de recomendação antes como parte de uma grande organização e trabalhei

através de todos os tipos de engenheiros e contabilidade para diferentes partes do problema.

É um dos que eu estou mais feliz porque, em última análise,

eu vim com uma solução muito simples que era fácil de entender de todos os níveis,

desde os executivos aos engenheiros e desenvolvedores.

Em última análise, era tão eficiente quanto algo realmente complexo,

e eles poderiam ter gasto muito mais tempo.

Na universidade,

temos um problema que queríamos prever as flores de algas.

Esta alga floresce pode causar um aumento na

toxicidade da água e pode causar problemas através da empresa de tratamento de água.

Não podíamos prever com nossa experiência em engenharia química.

Então usamos redes neurais artificiais para prever quando essas flores ocorrerão.

Assim, as empresas de tratamento de água poderiam lidar melhor com este problema.

Em Toronto, o transporte público é operado pela Comissão de Transit de Toronto.

Nós os chamamos de TTC. É uma

das maiores autoridades de trânsito da região, na América do Norte.

E um dia eles me contataram e disseram: “Temos um problema.”

E eu disse, “Ok, qual é o problema?”

Eles disseram: “Bem, temos dados de reclamações,

e gostaríamos de analisá-los, e precisamos da sua ajuda.”

Eu disse: “Tudo bem, eu ficaria muito feliz em ajudar.”

Então eu disse: “Quantas reclamações você tem?”

Eles disseram: “Alguns”. Eu disse:

“Quantos?” Talvez meio milhão.

Eu disse, “Bem, vamos começar a trabalhar com isso.”

Então peguei os dados e comecei a analisá-los.

Então, basicamente, eles fizeram um ótimo trabalho de manter

alguns dados em formato tabular que eram dados não estruturados.

E nesse caso, os dados tabulares foram quando a reclamação chegou,

quem a recebeu, qual era o tipo da reclamação,

foi resolvida, de quem foi a culpa.

E a parte não estruturada foi a troca de e-mails e faxes.

Imagine

como meio milhão de trocas de e-mails e tentando obter algumas respostas.

Então comecei a trabalhar com ele.

A primeira coisa que eu queria saber é por que as pessoas reclamariam

e há um padrão ou há alguns dias em que há mais queixas do que outras?

E eu olhei para os dados e os analisei em todos os formatos diferentes,

e não consegui encontrar o ímpeto

para as queixas serem maiores em um determinado dia e menores em outros.

E continuou por talvez um mês ou mais.

E então, um dia eu estava saindo do ônibus em Toronto,

e eu ainda estava pensando sobre isso.

E eu saí sem olhar para o chão,

e eu pisei em uma poça, poça de água.

E agora, eu estava meio que no tornozelo fundo da água,

e estava apenas um pé molhado e o outro seco.

E eu estava extremamente irritado.

E eu estava voltando e então ele me bateu,

e eu disse, “Bem, espere um segundo.

Hoje choveu inesperadamente,

e eu não estava preparado para isso.

É por isso que estou molhado, e eu não estava olhando para a frente.”

E se houvesse uma relação entre o

clima extremo e o tipo de reclamações que o TTC recebe?

Então eu fui para o ambiente do site do Canadá,

e eu tenho dados sobre chuva e precipitação,

vento e luz.

E lá, encontrei algo muito interessante.

Os 10 dias mais excessivos para reclamações.

Os 10 dias em que as pessoas mais reclamam foram os dias em que o tempo estava ruim.

Foi uma chuva inesperada,

uma queda extrema de temperatura,

muita neve, dia muito ventoso.

Então voltei para os executivos do TTC e disse:

“Tenho boas e más notícias.”

E a boa notícia é

que sei por que as pessoas reclamariam excessivamente em certos dias.

Eu sei a razão para isso. A má notícia é

que não há nada que possa fazer.

# Old problems, new problems, Data Science solutions

As organizações podem aproveitar a quantidade quase ilimitada de dados agora disponível para elas de um

número crescente de maneiras.

No entanto, todas as organizações usam a ciência de dados pelo mesmo motivo — para descobrir

soluções ideais para os problemas existentes.

Vamos dar uma olhada em três exemplos de ciência de dados fornecendo soluções inovadoras para

problemas antigos.

No transporte, a Uber coleta dados do usuário em tempo real para descobrir quantos motoristas estão disponíveis,

se forem necessários mais e se eles devem permitir uma sobrecarga para atrair mais motoristas. A

Uber usa dados para colocar o número certo de motoristas no lugar certo, na hora certa,

por um custo que o ciclista está disposto a pagar.

Em um esforço diferente de ciência de dados relacionados ao transporte, a Comissão de Transportes de Toronto

fez grandes progressos na resolução de um antigo problema com fluxos de tráfego, reestruturando esses fluxos

dentro e ao redor da cidade.

Usando ferramentas de ciência de dados e análise, eles têm:

Reuniram dados para entender melhor as operações de bonde e identificar áreas para intervenções

Analisaram dados de reclamações de clientes Dados de sonda usados para entender melhor o

desempenho do tráfego em rotas principais Criou uma equipe para melhor capitalizar

dados para planejamento, operações e avaliação

Ao focar nas autorizações de horário de pico e identificar as rotas mais congestionadas, as horas mensais perdidas

para os passageiros devido ao congestionamento do tráfego diminuíram de 4,75 horas em 2010 para 3 horas em meados de 2014.

Ao enfrentar problemas em nosso ambiente, a ciência de dados também pode desempenhar um papel proativo.

Os lagos de água doce fornecem uma variedade de necessidades humanas e ecológicas, tais como o fornecimento de

água potável e a produção de alimentos.

Mas os lagos em todo o mundo estão ameaçados por crescentes incidências de

flores nocivas de cianobactérias.

Existem muitos projetos e estudos para resolver este dilema há muito existente.

Nos EUA, uma equipe de cientistas de centros de pesquisa que se estendem do Maine à Carolina do Sul

está desenvolvendo e implantando ferramentas de alta tecnologia para explorar cianobactérias em lagos da

costa leste.

A equipe está usando barcos robóticos, bóias e drones equipados com câmeras para medir

dados físicos, químicos e biológicos em lagos onde cianobactérias são detectadas, coletando grandes

volumes de dados relacionados aos lagos e ao desenvolvimento das flores nocivas.

O projeto também está construindo novos modelos algorítmicos para avaliar as descobertas.

As informações coletadas conduzirão a melhores previsões de quando e onde

ocorrem as flores de cianobactérias, possibilitando abordagens proativas para proteger a saúde pública em lagos recreativos

e naqueles que fornecem água potável.

Esse treinamento interdisciplinar prepara a próxima geração de cientistas para abordar

questões sociais com as ferramentas adequadas de ciência de dados modernizadas.

É preciso reunir muitos dados, limpar e prepará-los e, em seguida, analisá-los para

obter a percepção necessária para desenvolver melhores soluções para as empresas de hoje.

Como obter uma solução melhor que seja eficiente?

Você deve: Identificar o problema e estabelecer uma

compreensão clara do mesmo.

Reúna os dados para análise.

Identifique as ferramentas certas a serem usadas.

Desenvolver uma estratégia de dados.

Estudos de caso também são úteis na personalização de uma solução potencial.

Quando essas condições existirem e os dados disponíveis forem extraídos, você poderá desenvolver um

modelo de aprendizado de máquina.

Levará tempo para uma organização refinar as práticas recomendadas para a estratégia de dados usando

ciência de dados, mas os benefícios valem a pena.

# Data Science Topics and Algorithms

Eu realmente gosto de regressão.

Eu diria que a regressão foi talvez um dos primeiros conceitos que eu, que realmente

me ajudou a entender os dados para que eu goste de regressão.

Eu realmente gosto de visualização de dados.

Eu acho que é um elemento-chave para as pessoas transmitirem sua mensagem para

pessoas que não entendem muito bem o que é a ciência de dados. Redes

neurais artificiais.

Eu sou realmente apaixonado por redes neurais porque nós temos muito a aprender com a natureza

então quando estamos tentando imitar nosso, nosso cérebro eu acho que podemos fazer algumas aplicações com

esse comportamento com esse comportamento biológico em algoritmos.

Visualização de dados com R. Eu adoro fazer isso.

Vizinho mais próximo.

É o mais simples, mas ele só obtém os melhores resultados muito mais vezes do que algum

algoritmo sobrecarregado, sobrecarregado que é tão provável de sobreajustar quanto é para fazer um bom ajuste.

Assim, os dados estruturados são mais como coisas de dados tabulares com as quais você está familiarizado no

formato Microsoft Excel.

Você tem linhas e colunas e isso é chamado de dados estruturados.

Dados não estruturados são basicamente dados que vêm principalmente da web onde não é

tabular.

Não é, não está em linhas e colunas.

É texto.

Às vezes é vídeo e áudio, então você teria que implantar algoritmos mais sofisticados

para extrair dados.

E, de fato, muitas vezes pegamos dados não estruturados e gastamos muito tempo e esforço

para obter alguma estrutura e depois analisá-los.

Então, se você tem algo que se encaixa bem em tabelas, colunas e linhas, vá de cabeça.

São os seus dados estruturados.

Mas se você vê se é um blog ou se você está tentando obter informações de páginas da Web

e você tem um zilhão de páginas da web, isso é dados não estruturados que exigiriam um

pouco mais de esforço para obter informações dele.

Há milhares de livros escritos sobre regressão e milhões de palestras entregues sobre regressão.

E eu sempre sinto que eles não fazem um bom trabalho de explicar a regressão porque eles

entram em dados, modelos e distribuições estatísticas.

Vamos esquecer isso.

Deixe-me explicar a regressão nos termos mais simples possível.

Se você já fez um passeio de táxi, um passeio de táxi, você entende a regressão.

É assim que funciona.

No momento em que você se senta em um táxi, em um táxi, você vê que há uma quantidade fixa lá.

Diz $2,50.

Você, preferindo o táxi, se move ou você sai.

Isto é o que você deve ao motorista no momento em que você entrar em um táxi.

Isso é uma constante.

Você tem que pagar essa quantia se você entrou em um táxi.

Então, à medida que começa a mover-se por cada metro ou cem metros, a tarifa aumenta em certa

quantidade.

Então há uma... há uma fração, há uma relação entre distância e o valor que

você pagaria acima e além dessa constante.

E se você não está se movendo e está preso no trânsito, então a cada minuto adicional você

tem que pagar mais.

Assim, à medida que os minutos aumentam, a sua tarifa aumenta.

À medida que a distância aumenta, sua tarifa aumenta.

E enquanto tudo isso está acontecendo você já pagou uma tarifa base que é a constante.

Isto é o que é a regressão.

Regressão informa qual é a tarifa base e qual é a relação entre

o tempo e a tarifa que você pagou, a distância que você viajou e a tarifa que você pagou.

Porque na ausência de conhecer esses relacionamentos, e apenas saber quanto as pessoas viajavam

e quanto pagavam, a regressão permite calcular essa constante que você não

sabia.

Que são $2,50, e calcularia a relação entre a tarifa e a distância,

a tarifa e o tempo.

Isso é regressão.

# Cloud for Data Science

A nuvem é uma dádiva de Deus para cientistas de dados.

Principalmente porque você é capaz de pegar seus dados,

pegar suas informações e colocá-las na nuvem,

colocá-las em um sistema de armazenamento central.

Ele permite que você ignore

as limitações físicas dos

computadores e dos sistemas que você está

usando e permite que você implante

as capacidades de análise e armazenamento

de máquinas avançadas que não precisam necessariamente

ser a sua máquina ou a máquina da sua empresa.

O Cloud permite que você não apenas armazene grandes quantidades de

dados em servidores em algum lugar na Califórnia ou em Nevada,

mas também permite que você implante

algoritmos de computação muito avançados e

a capacidade

de fazer computação de alto desempenho

usando máquinas que não são suas.

Pense nisso como você tem

alguma informação, você não pode armazená-la,

então você a envia para o espaço de armazenamento,

vamos chamá-lo de Nuvem,

e os algoritmos que você precisa usar,

você não os tem com você.

Mas então, na Nuvem,

você tem esses algoritmos disponíveis.

O que você faz é implantar esses algoritmos em

conjuntos de dados muito grandes e

você é capaz de fazê-lo mesmo que seus próprios sistemas,

suas próprias máquinas, seus próprios ambientes de computação

não estivessem permitindo que você o fizesse.

A nuvem é linda.

A outra coisa que a Cloud é

bonita para é que ela permite que

várias entidades trabalhem

com os mesmos dados ao mesmo tempo.

Você pode estar trabalhando com os mesmos dados

que seus colegas dizem

Alemanha e outra equipe na Índia

e outra equipe em Gana,

eles estão trabalhando coletivamente e

eles são capazes de fazê-lo porque as informações,

e os algoritmos, e as ferramentas,

e as respostas, resultados,

tudo o que eles precisavam está disponível em um local central,

que chamamos de Nuvem. A nuvem é linda.

O uso da nuvem permite que você obtenha acesso instantâneo a

tecnologias de código aberto, como o Apache Spark,

sem a necessidade de instalá-las e configurá-las localmente.

O uso da nuvem também oferece

acesso às ferramentas e

bibliotecas mais atualizadas sem a preocupação de

mantê-las e garantir que elas estejam atualizadas.

A Nuvem é acessível a partir de

todos os lugares e em cada fuso horário.

Pode utilizar tecnologias baseadas na nuvem

a partir do portátil, do tablet

e até do telemóvel,

permitindo a colaboração mais facilmente do que nunca.

Vários colaboradores ou equipes

podem acessar os dados simultaneamente,

trabalhando juntos na produção de uma solução.

Algumas grandes empresas de tecnologia oferecem plataformas de nuvem,

permitindo que você se familiarize com

tecnologias baseadas em nuvem em um ambiente pré-construído.

A IBM oferece a IBM Cloud, a

Amazon oferece a Amazon Web Services ou a AWS

e o Google oferece a plataforma Google Cloud. A

IBM também fornece laboratórios de Skills Network ou laboratórios de SN

para alunos registrados em qualquer um

dos portais de aprendizado da IBM Developer Skills Network,

onde você tem acesso a ferramentas

como Jupyter Notebooks e

clusters Spark para que você possa criar

seu próprio projeto de ciência de dados e desenvolver soluções.

Com prática e familiaridade,

você descobrirá como a nuvem

aumenta drasticamente a produtividade dos cientistas de dados.

Leitura

## Parabéns! Você foi aprovado!

Nota recebida 100%

Nota do envio mais recente 100%

Para ser aprovado 66% ou superior

Ir para o próximo item

### 1.

Pergunta 1

According to the reading, **the author** defines **data science** as the art of uncovering the hidden secrets in data.

1 / 1 ponto



True.



False.

Correto

Correct. The author defines data science as what data scientists do.

### 2.

Pergunta 2

What is admirable about **Dr. Patil’s** definition of a **data scientist** is that it limits data science to activities involving machine learning.

1 / 1 ponto



False.



True.

Correto

Correct. What is admirable about his definition is that it does not limit data science to activities involving machine learning.

### 3.

Pergunta 3

According to the reading, the characteristics exhibited by **the best** data scientists are those who are curious, ask good questions, and are O.K. dealing with unstructured situations.

1 / 1 ponto



True.



False.

Correto

Correct.

# Lesson Summary

In this lesson, you have learned:

* The typical work day for a Data Scientist varies depending on what type of project they are working on.
* Many algorithms are used to bring out insights from data.
* Accessing algorithms, tools, and data through the Cloud enables Data Scientists to stay up-to-date and collaborate easily.