Big Data e Data Science

I – Objetivo

"Nós estamos testemunhando um movimento que irá transformar completamente qualquer negócio e a sociedade. O nome que nós damos a esse movimento é **Big Data** e irá mudar tudo, a maneira que banco e varejistas operam, a forma que tratamos o câncer e protegemos o mundo contra o terrorismo. Não importa qual o trabalho que você está fazendo ou a indústria que você trabalha, Big Data irá transformá-lo" (Bernard Marr, 2016).

Big Data, basicamente, refere-se ao fato de que agora nós podemos coletar e analisar dados de formas que eram inimagináveis há alguns anos atrás. E há duas coisas que impulsionam esse movimento: o fato que temos mais dados de tudo e a nossa habilidade de guardar e analisar qualquer dado. Para exemplificar, estima-se que nós criamos mais dados nos últimos dois anos que em toda a nossa história. O avanço da computação distribuída significou que um grande volume de dados pudesse ser armazenado (em pequenos pedaços através de várias bases de dados) e analisados compartilhando as análises em diferentes servidores (cada um executando uma pequena parte da análise).

A quantidade de empresas que utilizam "**Big Data e Data Science**" para buscar soluções é extenso. Por exemplo, os *Data Scientists* do Walmart, maior rede varejista do mundo, em um Halloween perceberam, através do monitoramento das vendas e análise dos dados, que um novo *cookie* (é um biscoito e não um pequeno pacote de dados enviados de um website para o navegador do usuário quando o usuário visita um site) estava com as vendas abaixo do esperado em algumas lojas. Tal fato, fez com que o laboratório de dados acionasse o departamento de vendas, que descobriram que nessas lojas os *cookies* não haviam sido colocados na prateleira.

Esse curso, além de abordar o estado da arte sobre Big Data (e.g. Ecossistema Hadoop, Hive, Pig, NoSQL), utilizará para a análise de dados o software preferido da maior comunidade de cientistas de dados do mundo (Kaggle - http://goo.gl/j7b19s), o **software R**. Com mais de 2 milhões de usuários mundo afora, o **R** está se tornando rapidamente a linguagem de programação líder em *Data Science* e Estatística. Todo ano o número de usuários cresce a taxa de 40% e um grande número de empresas estão usando o R em suas atividades do dia a dia.

Ao terminar o curso você estará apto a trabalhar com grandes bases de dados, conhecerá o Python (outro software muito utilizado por cientistas de dados) e terá profundo conhecimento sobre o R. Você entenderá Estatística e Machine Learning. Saberá visualizar dados e criar Relatórios Dinâmicos. Compreenderá a magia por trás da Inteligência Artificial. Saberá como prever sua receita e qual é o impacto que as mídias sociais poderão causar sobre ela. Analisará espacialmente os dados e entenderá como a Big Data está influenciando os modelos financeiros.

Então, ao invés de colocar sua cabeça na areia ou ficar perdido nesse novo mundo chamado "**Big Data e Data Science**" você deveria se inscrever nesse curso e encontrar maneiras inteligentes de criar valor com as informações que estão por aí.

II - Conteúdo Programático

a) **Big Data e Data Science + Introdução ao R**

- **a. O que é Data Science?**
- **b. O que é Big Data?**
- **c. Roadmap para se tornar um cientista de dados**
- **d. R software**
  - i. Conceitos básicos e a filosofia do R
  - ii. Conhecendo o Ambiente (R e RStudio)
  - iii. Diferentes tipos de variáveis
  - iv. Objetos (Vetor, Data Frames, Matriz)
  - v. Trabalhando com listas
  - vi. Estrutura de condição: If, else e ifelse
  - vii. Estrutura de Repetição
  - viii. Funções
  - ix. Leitura\exportação arquivos (.xlsx,.csv,.txt entre outros)

b) **Estatistica**

- **a. Modelos probabilísticos e modelos estatísticos**
- **b. Revisão de Probabilidades:**
  - i. Distribuições de Probabilidades discretas e contínuas
  - ii. Esperança e Variância.
  - iii. Funções do R: dxxx, pxxx, qxxx, rxxx.
  - iv. Lei dos Grandes Números e Teorema Central do Limite
- **c. Inferência Estatística:**
  - i. Estimação Pontual
  - ii. Estimação por Intervalos
  - iii. Testes de Hipóteses
  - iv. Modelos Bayesianos
- d. **Modelagem**:
  - i. Modelos de Regressão Linear
  - ii. Modelos de Regressão Logística
- **e. Análise de dados amostrais complexos**
  - i. Pesquisa Nacional de Amostra de Domicílios Contínua (PNADC)
- **f. Modelagem Bayesiana: library rstan**

c) **Data Management e Computação na Nuvem**

- **a. Introdução a Bancos de Dados Estruturados**
- **b. Diagramas Entidade-Relacionamento**
- **c. SQL**
  - i. Criação de bancos e tabelas
  - ii. Leitura de bases de dados grandes
  - iii. Comandos de consulta
- **d. Utilização do MySQL**
- **e. Conceitos básicos de Cloud**
  - i. Características
  - ii. Benefícios

II - Conteúdo Programático

iii. Riscos

f. Modelos de Serviço na Nuvem

i. Software as a Service
ii. Platform as a Service
iii. Infrastructure as a Service

g. Serviços na Nuvem

h. Utilização do Microsoft Azure

i. Data mining com R

d) Visualização de dados e Dynamic reports

a. Gráficos:

i. Pontos
ii. Barras
iii. Pizza e Diagrama de Venn
iv. Histograma e Boxplot
v. Grafos
vi. Matriz de correlação
vii. Mapa de árvore (Tree Map)
viii. Nuvem de palavras (Word Cloud)
ix. Linha (para séries temporais)

b. Relatórios dinâmicos e visualização de dados fora do R

i. R Markdown
ii. Shiny

e) **Bancos de Dados em Larga Escala: Hadoop e NoSQL**

a. **Fundamentos de Hadoop**

i. Surgimento
ii. Objetivo
iii. Arquitetura
iv. Hadoop 1 X Hadoop 2
v. Distribuições
vi. Administração
vii. Ecossistema Hadoop
viii. Níveis de Maturidade em Análise de Dados
ix. Business Analytics X Business Intelligence
x. Data Lake
xi. Hands On sobre Fundamentos

b. **Hive**

i. Introdução ao Hive
ii. Hands On:
I. Importar e exportar dados,
II. Criação de bancos e tabelas,
III. Operações básicas

c. **Pig**

i. Introdução ao Pig
ii. Hands on:
I. Como ler dados (READ),
II. Como escrever dados (OUTPUT),

II - Conteúdo Programático

III. Operadores,
IV. Funções.

d. **Introdução a NoSQL**
   i. Conceitos e características
   ii. Teorema de CAP
   iii. Tipos de bancos NoSQL
   iv. Casos de Uso

e. **Stack ELK**
   i. Introdução;
   ii. Arquitetura;
   iii. Logstash;
   iv. ElasticSearch;
   v. Kibana;
   vi. Hands On

f) Machine Learning

a. Introduction to Machine Learning
   i. Exemplos de utilização
   ii. Motivos para Estimar
   iii. Como estimar
   iv. Trade-off precisão-interpretabilidade
   v. Aprendizado Supervisionado e Não Supervisionado
   vi. Trade-Off Vício-Variância

b. Linear Regression
   i. Representação do modelo e função custo
   ii. Estimação de coeficientes
   iii. Gradient Descent

c. Classification
   i. Logistic Regression

d. Resampling Methods
   i. Cross-Validation

e. Regularization
   i. Shrinkage Methods (Ridge Regression e Lasso)
   ii. Dimension Reduction Methods
   iii. Problemas da dimensionalidade

f. Métodos Baseados em Árvores
   i. Regression Trees
   ii. Classification Trees
   iii. Bagging
   iv. Random Forests
   v. Boosting

g. Support Vector Machines
   i. Optimization objective
   ii. Large Margin intuition

h. Unsupervised Learning
   i. PCA
   ii. K-Means CLustering
   iii. Hierarchical Clustering

II - Conteúdo Programático

i. **Machine Learning at Scale**
   i. Gradient descent at scale
   ii. Online Learning
   iii. Parallelism

g) **Estatística Espacial**

a. **Visão Geral**
   i. O que é Análise Espacial
   ii. Tipos de processos espaciais
   iii. Conceitos Gerais
   iv. Sistema de Informações Geográficas (GIS)

b. **Processos pontuais espaciais**
   i. Mapas interativos no R
   ii. Identificação de dependência espacial
   iii. Processo de Poisson

c. **Dados de área**
   i. Visualização e análise exploratória
   ii. Principais modelos: CAR e SAR

d. **Geoestatística**
   i. Interpolação espacial
   ii. Regressão espacial
   iii. Previsão linear

e. **Análise espacial de cluster**

h) **Análise de Séries Temporais**

a. **Modelos (S)ARIMA**
   i. Processos Auto-Regressivos de Médias Móveis – ARMA(p,q)
   ii. Identificação
   iii. Estimação
   iv. Diagnóstico dos Resíduos
   v. Previsão
   vi. Modelagem da série temporal de venda de passagens aéreas (*AirPassengers*)
   vii. *Hands-on*: Previsão da Produção Industrial - PIM-PF (IBGE) usando o R

b. **Modelos de Regressão Dinâmica**
   i. Modelo clássico de regressão linear
   ii. O problema da correlação serial
   iii. Modelos autoregressivos com defasagens distribuídas (*Autoregressive Distributed Lag* (ADL)
   iv. O problema da Cointegração e o Mecanismo de Correção de Erros (ECM)
   v. Regressão Espúria
   vi. The dunk and her dog
   vii. Teste de Cointegração de Engle-Granger

II - Conteúdo Programático

viii. Modelos de Correção de Erros

c. Automação de rotinas de captura de dados;

d. Conhecendo os principais bancos de dados;

i. FRED package
ii. Anbima;
iii. Banco Central;
iv. Brazilian Economic Time Series (BETS) package

i) Redes Neurais e Aplicações

a. Redes Neurais Artificiais

i. Breve histórico do desenvolvimento das redes neurais artificiais
ii. Estrutura do neurônio artificial
iii. Perceptron
iv. Regra de Hebb
v. O problema do OU-Exclusivo
vi. Regra delta
vii. Multilayer Perceptron com Backpropagation
viii. Previsão de Séries Temporais
ix. *Hands-on*: Previsão de séries temporais com redes neurais usando o R
x. Classificadores Bayesianos Robustos
xi. *Hands-on*: Classificação com o R

b. Algoritmos genéticos

i. Seleção natural e evolução
ii. Componentes de um AG tradicional
iii. Operadores genéticos (Reprodução, Seleção, Mutação e Crossover)
iv. Fundamentos matemáticos
v. Teoria de Schema
vi. *Hands-on*: Ajuste dos hiperparâmetros de modelos de alisamento exponencial usando o R

j) Mídias Sociais

a. Análise textual utilizando o R

i. Análise descritiva
ii. Análise de sentimento utilizando dicionário
iii. Métodos supervisionados
iv. Métodos não supervisionados

b. Webscraping

i. Dados Estruturados
ii. Dados não estruturados
iii. APIs

c. Coleta e análise de dados de mídias sociais

i. Facebook
ii. Twitter

II - Conteúdo Programático

iii. Estudo de caso

k) Big Data Analytics pela ótica de negócios

a. Qual o real ritmo da mudança?

i. Introdução da evolução da Tecnologia de armazenamento de dados

b. Analytics: por que agora?

i. Qual o tamanho da explosão de dados / desafios e oportunidades
ii. As empresas brasileiras estão reconhecendo a necessidade de ter mais inteligência e insights sobre seus negócios

c. Big Data Analytics pela ótica de negócios

i. O dado é considerado o quarto fator de produção
ii. Como melhorar o desempenho empresarial
iii. Abordagem tradicional versus abordagem direcionada pelo negócio
iv. Ciclo de vida de um projeto de Analytics

d. Panorama do mercado e principais players

e. Cases Reais de Advanced Analytics

i. Análise preditiva de Produtividade de florestal
ii. Previsão de audiência
iii. Governança Jurídica com Big Data Analytics
iv. Avaliação e valoração de impacto socioambiental com aplicação de Big Data Analytics
v. Como o analytics ajuda o Comitê Olímpico Brasileiro tomar melhores decisões nos investimentos nos esportes olímpicos
vi. Computação cognitiva

f. Procurando por um ponto de partida para iniciar um projeto de analytics?

i. Como utilizar as fontes de dados disponíveis em minha corporação para solucionar problemas atuais, respeitando minha estratégia corporativa, ao mesmo tempo que me aproxima de minha missão e visão?
ii. Workshop – hands-on
iii. Diagrama Causal
iv. Descoberta de fonte de informações

Coordenador: Pedro Costa Ferreira

Doutor em Engenharia Elétrica - (*Decision Support Methods*) e Mestre em Economia. Co-autor dos livros "Planejamento da Operação de Sistemas Hidrotérmicos no Brasil" e "Análise de Séries Temporais em R: um curso introdutório". É o primeiro pesquisador da América Latina a ser recomendado pela empresa RStudio Inc. Atuou em projetos de Pesquisa e Desenvolvimento (P&D) no setor elétrico nas empresas Light S.A. (e.g. estudo de contingências judiciais), Cemig S.A, Duke Energy S.A, entre outras. Ministrou cursos de estatística e séries temporais na PUC-Rio e IBMEC e em empresas como o Operador Nacional do Setor Elétrico (ONS), Petrobras e CPFL S.A. Atualmente é professor de Econometria de Séries Temporais e Estatística e cientista chefe do Núcleo de Métodos Estatísticos e Computacionais (FGV|IBRE). É também revisor de importantes journals, como Energy Policy e Journal of Applied Statistics. Principais estudos são em modelos Econométricos, Setor Elétrico, Incerteza Econômica, Preços, R software e Business Cycle.

email: pedro.guilherme@fgv.br
GitHub: https://github.com/pedrocostaferreira
Website: https://pedrocostaferreira.github.io/

AULA A AULA			
Dia/Mês	**Conteúdo**	**Tópico**	**Leitura Prévia**
Aula 0	**Definição Data Science** **Definição Big Data** **Roadmap para se tornar um cientista de dados** **R software:**Conceitos básicos e a filosofia do R;Conhecendo o Ambiente (R e RStudio); Diferentes tipos de variáveis; Objetos (Vetor, Data Frames, Matriz); Trabalhando com listas; Estrutura de condição: If, else e ifelse; Estrutura de Repetição; Funções; Leitura\exportação arquivos (.xlsx,.csv,.txt entre outros)	**Nivelamento**	
1ª aula	**Modelos de Probabilidade:** Distribuições discretas e contínuas; Medidas de Centro e de Dispersão; Funções do R para calcular probabilidades, percentis e gerar dados de várias distribuições. Resultados de limites evidenciados por meio de simulação no R **Modelos Estatísticos:** Famílias de distribuições paramétricas; Inferência sobre parâmetros; Problema de medição. **Estimação Pontual**: Estimadores não-tendenciosos; Estimação de Máxima Verossimilhança. Estimadores consistentes	**Big Data e Data Science + Introdução ao R**	
2ª aula	**Modelos de Probabilidade:** Distribuições discretas e contínuas; Medidas de Centro e de Dispersão; Funções do R para calcular probabilidades, percentis e gerar dados de várias distribuições. Resultados de limites evidenciados por meio de simulação no R **Modelos Estatísticos:** Famílias de distribuições paramétricas; Inferência sobre parâmetros; Problema de medição. **Estimação Pontual**: Estimadores não-tendenciosos; Estimação de Máxima Verossimilhança. Estimadores consistentes	**Estatística**	

AULA A AULA			
Dia/Mês	**Conteúdo**	**Tópico**	**Leitura Prévia**
3ª aula	**Intervalos de Confiança** – Nível de confiança e intervalos aproximados **Teste de Hipótese**: Nível de significância, Erros de Tipos I e II, Teste-Z ; Teste-t. Relação entre intervalo de confiança e teste de hipótese **Modelos Bayesianos**. **Modelos Lineares**: Regressão linear simples; Teste-t e teste-F.	**Estatística**	
4ª aula	**Modelos Lineares:** Regressão linear múltipla, testes, variáveis explicativas categóricas, interação; seleção de variáveis. **Regressão Logística:** Ajuste de modelos e uso em classificação. **Análise de dados Amostrais complexos**: library survey do R; Exemplo da PNADC. **Stan** – library rstan do R. Exemplo simples de utilização	**Estatística**	
5ª aula	Introdução a bancos de dados estruturados. Diagramas ER. SQL: Comandos de criação, manutenção e consultas. MySQL: Carregamento de grandes bases de dados. Utilização do MySQL Workbench. Integração com o R.	**Data Management e Computação na Nuvem**	
6ª aula	Conceitos basicos de computacao na nuvem. Tipos e modelos de servicos na nuvem. Utilizacao do Microsoft Azure. Data Mining com o R.	**Data Management e Computação na Nuvem**	
7ª aula	**Gráficos:** Pontos; Barras; Pizza e Diagrama de Venn; Histograma e Boxplot; Grafos; Matriz de correlação; Mapa de árvore (Tree Map); Nuvem de palavras (Word Cloud); Linha (para séries temporais). R Markdown: Instalação; Gerando documentos dinâmicos; Publicando na web.	**Visualização de dados e Dynamic reports**	
8ª aula	**Shiny:** Instalação; Desenvolvendo aplicativos básicos; Lendo base de dados locais; Adicionando imagens e documentos ao Shiny; Personalizando o Shiny; Publicando aplicativos na web.	**Visualização de dados e Dynamic reports**	

AULA A AULA			
Dia/Mês	**Conteúdo**	**Tópico**	**Leitura Prévia**
9ª aula	**Big Data:** Surgimento, 3V's, Escalabilidade Vertical X Escalabilidade Horizontal **Hadoop:** Surgimento, Conceitos, Arquitetura Hadoop 1 X Hadoop 2, Ecossistema Hadoop. **BI:** KDD, BI x Big Data , Data Lake, Níveis de Maturidade.	**Hadoop e NoSQL**	
10ª aula	**Ambari:** Conceitos de monitoramento e manipulação do HDFS via interface. HDFS: Conceitos e manipulação de arquivos via console. **Hive:** Conceitos, Arquitetura, Funções, Integração com R, Integração com Tableau.	**Hadoop e NoSQL**	
11ª aula	**Pig:** Conceitos, Pig X Hive, Arquitetura e funções. **NoSQL:** Conceitos, ACID x BASE, Teorema de CAP, Tipos de Banco NoSQL. **ELK**: Conceitos, Arquitetura, Sharding X Replica, Case Real Time.	**Hadoop e NoSQL**	
12ª aula	**Introduction to Machine Learning:** Exemplos de utilização e principais trade-offs **Linear Regression:** Métodos numéricos de estimação **Classification:** Regressão Logística **Resampling Methods:** Cross-Validation **Regularization:** Ridge Regression, Lasso e o Problema da dimensionalidade	**Machine Learning**	
13ª aula	**Métodos Baseados em Árvores:** Trees, Bagging, Random Forests e Boosting **Support Vector Machines:** Optimization objective e Large Margin intuition **Unsupervised Learning:** PCA, K-Means CLustering e Hierarchical Clustering **Machine Learning at Scale:** Gradient descent at scale, Online Learning e Parallelism	**Machine Learning**	

AULA A AULA			
Dia/Mês	**Conteúdo**	**Tópico**	**Leitura Prévia**
14ª aula	**Introdução à análise espacial:** dados espaciais x dados não espaciais; conceitos gerais; sistema de informações geográficas; **Processos pontuais espaciais:** mapas interativos no R; estimação via kernel; distâncias para vizinho mais próximo (função F e G); função K; aleatoriedade espacial completa; processos de Poisson. **Dados de área:** visualização de dados de área; análise exploratória.	**Estatística Espacial**	
15ª aula	**Dados de área:** índice de Moran; índice de Geary; indicadores locais de associação espacial (LISA); modelos CAR e SAR. **Geoestatística:** visualização interativa no R; interpolação espacial; modelo de regressão espacial; krigagem. **Análise espacial de cluster:** K-means; CLARA; AGNES; DIANA; DBSCAN.	**Estatistica Espacial**	
16ª aula	**Modelos Univariados:** Modelos ARIMA; **Hands-on:** Previsão da Produção Industrial - PIM-PF (IBGE) usando o R	**Séries Temporais e Modelos Econométricos**	Ferreira et. al.(capítulo 5)
17ª aula	**Modelos de Regressão Dinâmica:** Modelo clássico de regressão linear; O problema da correlação serial; Modelos autoregressivos com defasagens distribuídas (Autoregressive Distributed Lag (ADL); Regressão Espúria; The dunk and her dog; Teste de Cointegração de Engle-Granger	**Séries Temporais e Modelos Econométricos**	Ferreira et. al.(capítulo 8)
18ª aula	**Redes Neurais Artificiais**: Neurônio artificial, perceptron, Regra delta, Multilayer Perceptron com Backpropagation, Previsão de Séries Temporais, Classificadores Bayesianos.	**Inteligência Artificial**	
19ª aula	**Lógica Fuzzy:** Sistema de inferência fuzzy, Wang e Mendel, Previsão de séries temporais, **Algoritmo Genético**	**Inteligência Artificial**	

AULA A AULA			
Dia/Mês	**Conteúdo**	**Tópico**	**Leitura Prévia**
20ª aula	**Análise textual utilizando o R:** Análise descritiva, Análise de sentimento utilizando dicionário, Métodos supervisionados e Métodos não supervisionados **Webscraping**: Dados Estruturados, Dados não estruturados e APIs	**Mídias Sociais**	
21ª aula	**Coleta e análise de dados de mídias sociais:** Facebook, Twitter e Estudo de caso	**Mídias Sociais**	
22ª aula	Qual o real ritmo da mudança? Analytics: por que agora? Big Data Analytics pela ótica de negócios	**Big Data Analytics pela ótica de negócios**	
23ª aula	Panorama do mercado e principais players; Cases Reais de Advanced Analytics; Procurando por um ponto de partida para iniciar um projeto de analytics?	**Big Data Analytics pela ótica de negócios**	