

Programa de Especialização em Data Science e Big Data

Contextualização e Estrutura do Curso

Prof. Wagner Hugo Bonat

Curso de Especialização em
Data Science & Big Data
Universidade Federal do Paraná

2 de março de 2018



Mercado de trabalho em data science e analytics (DSA)

- ▶ The quant crunch: How the demand for data science skills is disrupting the job marketing. IBM, 2017.
 - ▶ Projeções para o horizonte 2020.
 - ▶ 364.000 novas vagas para cientistas de dados e similares serão abertas
 - ▶ Demanda por cientistas e engenheiros de dados deve crescer 39%.
 - ▶ Vagas para DSA ficam abertas em média 5 dias a mais.
 - ▶ Salário anual médio \$ 80,265.
 - ▶ Crescimento estimado em \$187 billion.

Por que este crescimento ?

- ▶ Tudo que fazemos hoje em dia envolve troca/geração de dados.
- ▶ Aplicativos para *smartphones* e computadores pessoais são coletores de dados poderosos.
- ▶ Todo dia 2.5 quintillion bytes de dados são criados (IDC - International Data Corporation).
- ▶ Dados estão nas mais diversas formas (mapas, fotos, videos, audios etc).
- ▶ Entender o comportamento das pessoas é uma ferramenta de negócio extremamente poderosa.
- ▶ Entender/analisar esta massa de dados e tirar “proveito” deles é o trabalho do cientista de dados.

Cientista de dados

- ▶ Profissão extremamente nova e definições ainda estão em formulação.
- ▶ A *Data Science Association* definiu em 2013 os termos ciência de dados e cientista de dados em seu *Data Science Code of Professional Conduct* (www.datascienceassn.org).
 - ▶ Ciência de dados: é o estudo científico da criação, validação e transformação de dados para criar significado.
 - ▶ Cientista de dados: é o profissional que usa métodos científicos para entender e criar significado a partir de dados brutos.

“Person who is better at statistics than any software engineer and better at software engineering than any statistician - Josh Wills.”

O trabalho do Cientista de dados

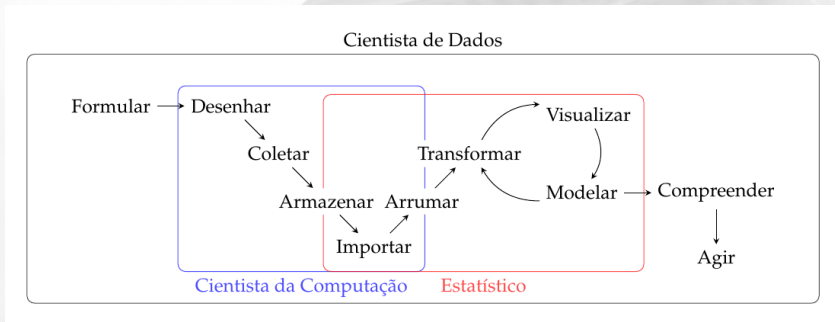
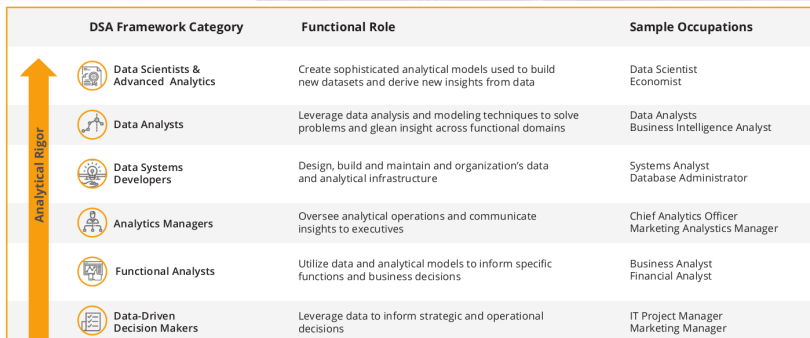


Figura 1: Workflow típico de um Cientista de Dados.

A profissão no mercado de trabalho

- O profissional em DSA aparece com diversas categorias.









DSA Framework Category	Functional Role	Sample Occupations
 Data Scientists & Advanced Analytics	Create sophisticated analytical models used to build new datasets and derive new insights from data	Data Scientist Economist
 Data Analysts	Leverage data analysis and modeling techniques to solve problems and glean insight across functional domains	Data Analysts Business Intelligence Analyst
 Data Systems Developers	Design, build and maintain an organization's data and analytical infrastructure	Systems Analyst Database Administrator
 Analytics Managers	Oversee analytical operations and communicate insights to executives	Chief Analytics Officer Marketing Analytics Manager
 Functional Analysts	Utilize data and analytical models to inform specific functions and business decisions	Business Analyst Financial Analyst
 Data-Driven Decision Makers	Leverage data to inform strategic and operational decisions	IT Project Manager Marketing Manager

Figura 2: Descrição de DSA cargos - IBM 2017.

Habilidades analíticas por DSA categoria.

DSA Framework Category	Occupation	Analytical Score (2015)
Analytics Managers	Financial Analysis SQL SAS Data Analysis Business Intelligence	Budgeting Project Management Risk Management Accounting Financial Planning
Data Analysts	Data Analysis SQL Business Intelligence Data Warehousing SAS	Project Management Microsoft Access Business Process SAP Business Analysis
Data Systems Developers	SQL Database Administration Extraction, Transformation, and Loading Data Warehousing Apache Hadoop	Project Management LINUX Software Development UNIX JAVA
Data Scientists & Advanced Analysts	Apache Hadoop Machine Learning Big Data R Data Science	Python JAVA Economics C++ Project Management
Data-Driven Decision Makers	SQL Financial Analysis Data Analysis Data Management Data Validation	Budgeting Project Management Accounting Supervision Product Management
Functional Analysts	Financial Analysis SQL Data Analysis Data Management SAS	Budgeting Accounting Business Analysis Business Process Economics

Figura 3: Habilidades por DSA cargo - IBM 2017.


Habilidades mais procuradas em DSA.

Skill Name	Total Postings in 2015
SQL	338,555
Data Analysis	166,285
Financial Analysis	155,331
Data Management	113,807
Mathematics	107,297
Data Warehousing	97,797
SQL Server	93,630
Database Administration	92,256
Business Intelligence	88,603
Extraction, Transformation, and Loading (ETL)	82,920

Figura 4: Habilidades mais procuradas em DSA - IBM 2017.

Habilidades que mais vão crescer em DSA.

Skill Name	Predicted 2-Year Growth
Data Science	93%
Machine Learning	56%
Tableau	52%
Big Data	50%
Data Visualization	44%
R	41%
Apache Hive	41%
Predictive Analytics	39%
Apache Hadoop	35%
Pivot Tables	34%

Figura 5: Habilidades que mais vão crescer em DSA - IBM 2017. 

Habilidades chaves para os cargos com maiores salários.

Occupation	Key Skills	High-Paying Skills
Data Scientist	Data Science Machine Learning Python R Apache Hadoop	Pattern Recognition Database Schemas Quantitative Analysis Object-Oriented Analysis and Design Database Administration
Data Engineer	Data Engineering Big Data Apache Hadoop JAVA Python	Spark Programming Oozie Predictive Models Apache Flume PIG
Finance and Risk Analytics Manager	Risk Management Financial Analysis and Planning Forecasting and Financial Modeling Project Management SQL	MATLAB Mergers and Acquisitions Data Warehousing Project Management R

Figura 6: Habilidades chaves para maiores salários - IBM 2017.

Matriz de DSA habilidades.

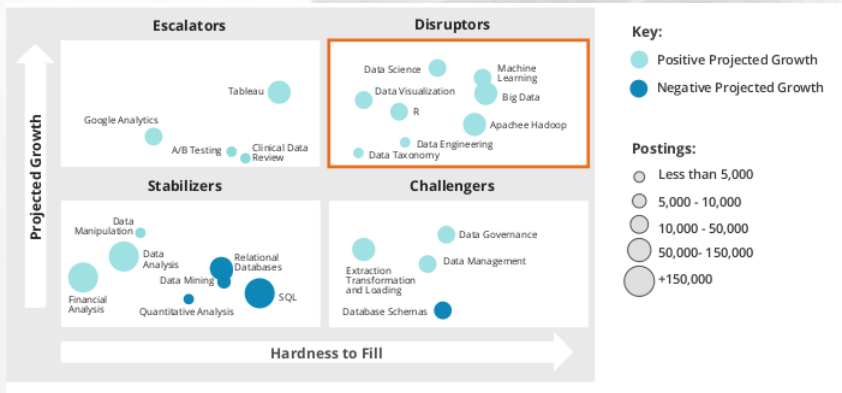


Figura 7: Matriz de DSA habilidades - IBM 2017.

Situação do Brasil

- ▶ Bacharelado em Estatística.
- ▶ Bacharelado em Ciência da Computação.
- ▶ Cientista de Dados unifica as áreas com uma visão integrada.
- ▶ Não há cursos de graduação no Brasil.
- ▶ Existe um grande espaço de oportunidades entre as áreas.
- ▶ Oportunidade de especializar em uma área de concentração.

Programa de Especialização em Data Science e Big Data

Objetivos

- ▶ Formar profissionais capazes de aplicar o estado da arte em termos de sistemas para Big Data e ferramentas de análise, visualização e apresentação de bases de dados com foco no desenvolvimento de soluções para negócios.

Perfil do egresso

- ▶ Profissional especializado em extrair conhecimento de grande volumes de dados usando tecnologias computacionais, análise estatística e conhecimento de negócios.

Habilidades chaves do Cientista de dados.

1. Programação/Banco de dados.

- 1.1 Fundamentos da ciência da computação.
- 1.2 Linguagens de programação R e Python.
- 1.3 Estatística computacional.
- 1.4 Banco de dados SQL e NoSQL.
- 1.5 Computação paralela.
- 1.6 Conceitos de MapReduce.
- 1.7 Hadoop e Hive/Pig.

2. Matemática/Estatística.

- 2.1 Modelagem estatística.
- 2.2 Delineamento de experimentos e amostragem.
- 2.3 Inferência Bayesiana e métodos computacionalmente intensivos.
- 2.4 Machine learning (supervisionada e não supervisionada).

3. Comunicação e Visualização.

- 3.1 Traduzir resultados da análise para não especialista.
- 3.2 Ferramentas de visualização como ggplot2, lattice, D3.js.

Estrutura do curso

6 módulos de 60 hrs + 1 módulo de 30 hrs.

1. Núcleo básico:

- 1.1 Infraestrutura computacional (60 hrs).
- 1.2 Inferência estatística para ciência de dados (60 hrs).
- 1.3 Linguagens de programação para ciência de dados (60 hrs).

2. Núcleo avançado:

- 2.1 Processamento de Big Data (60 hrs).
- 2.2 Modelos estatísticos (60 hrs).
- 2.3 Mineração de dados e aprendizado de máquina (60 hrs)
- 2.4 Métodos de pesquisa (30 hrs).

Infraestrutura computacional

1. Infraestrutura de software.

- 1.1 Shell scripting para automação de tarefas.
- 1.2 Estrutura de sistemas operacionais.

2. Infraestrutura de comunicação.

- 2.1 Conceitos e aplicações de redes.
- 2.2 Protocolos web.
- 2.3 Nuvem computacional.

3. Infraestrutura de alto desempenho.

- 3.1 Paralelismo e distribuição.
- 3.2 Clusters computacionais.
- 3.3 Submissão e controle de tarefas em clusters.

- Dr. Daniel Weingaertner.
- Dr. Luis Carlos Erpen de Bona.
- Dr. Marco Antonio Zanata Alves.

Inferência estatística para ciência de dados

1. Introdução à probabilidade e variáveis aleatórias.
 - 1.1 Distribuições discretas e contínuas de probabilidade.
 - 1.2 Esperança, variância e covariância.
 - 1.3 Resultados assintóticos (lei dos grandes números e teorema central do limite).
 2. Noções de amostragem e tipos de estudos amostrais.
 3. Paradigmas e procedimentos/elementos de inferência estatística.
 - 3.1 Estimação e funções de evidência.
 - 3.2 Intervalos de confiança e testes de hipóteses.
 - 3.3 Inferência baseada em simulação.
-
- ▶ Dr. Paulo Justiniano Ribeiro Jr.
 - ▶ Dr. Wagner Hugo Bonat.
 - ▶ Ms. Eduardo Vargas Ferreira (Mestre).

Linguagens de programação para ciência de dados

1. Introdução e configuração do ambiente de programação.

- 1.1 Operações aritméticas e lógicas.
- 1.2 Estruturas de controle e repetições.
- 1.3 Programação orientada à objetos e funções.

2. Leitura e manipulação de dados.

- 2.1 Importação e manipulação de dados em formato tabular.
- 2.2 Conexão e manipulação com sistemas de banco de dados.
- 2.3 Aquisição de dados da internet (web scraping).

3. Análise exploratória e visualização interativa de dados.

- ▶ Dr. André Abed Grégio.
- ▶ Dr. Elias Teixeira Krainski.
- ▶ Dr. Fernando de Pol Mayer.
- ▶ Dr. Walmes Marques Zeviani.

Processamento de Big Data

1. Introdução ao BigData.

1.1 Organização física e lógica.

1.2 Métodos de acesso.

2. Armazenamento de BigData.

2.1 SQL, NoSQL e NewSQL.

2.2 Data Warehouse.

2.3 Plataformas distribuídas.

3. Interfaces de visualização e interação.

► Dr. Eduardo Cunha de Almeida.

Modelos estatísticos

1. Modelos lineares.
2. Modelos lineares generalizados.
3. Seleção de modelos e penalização.
4. Modelos aditivos generalizados.
5. Árvores de regressão e classificação.
6. Regressão multivariada.
7. Imputação de dados.
8. Tópicos adicionais (sob demanda).

- ▶ Dr. César Augusto Taconeli.
- ▶ Dr. José Luiz Padilha da Silva.
- ▶ Dr. Elias Teixeira Krainski.

Mineração de dados e aprendizagem de máquina

1. Introdução ao Machine Learning.
 2. Representação de dados e engenharia de características.
 3. Aprendizagem supervisionada.
 4. Aprendizagem não supervisionada.
 5. Redução de dados/dimensão.
 6. Avaliação e melhoria de modelos.
 7. Algoritmos em cadeia e pipelines.
 8. Mineração de texto.
- Dr. Luiz Eduardo Soares de Oliveira.
 - Dr. Walmes Marques Zeviani.
 - Ms. Eduardo Vargas Ferreira.

Métodos de Pesquisa

1. Tipos de estudos e o método científico.
 2. Estrutura do artigo científico.
 3. Métodos e práticas da pesquisa reproduzível.
 4. Introdução às técnicas de apresentação e redação científica.
 5. Técnicas e ferramentas para elaboração de relatórios dinâmicos.
- ▶ Dr. Fernando de Pol Mayer.
 - ▶ Dr. Marco Antonio Zanata Alves.
 - ▶ Dr. Wagner Hugo Bonat.

Filosofia de ensino e avaliação

Curso será baseado em estudos de casos reais com muitas atividades práticas. Avaliação prioritariamente na forma de trabalhos práticos resolvendo problemas reais. Recomendado uma avaliação a cada 20 hrs/aula.

Certificação

► Requisitos básicos

- 75% de presença (13 finais de semana).
- 70% de aproveitamento nas disciplinas.
- Defesa da monografia.
- Banca composta de dois professores.

► Monografia

- Artigo para conferência (8-10 pgs).
- Assunto a ser definido com o orientador.
- Apresentação poster no evento do curso.

► Orientação

- Cada orientador vai orientar até dois alunos por turma.
- Co-orientação a critério dos professores e alunos.

Informações complementares

► Aulas

- Carga horária de 12 horas semanais em 3 módulos de 4 horas.
- Sextas-feiras das 19h às 23h.
- Sábados das 8h às 12h e 13h30 às 17h30.
- Aulas no 1o e 2o semestres (34 finais de semana em 2018).
- Monografia no 3o semestre (6 meses de 2019).

► Datas

Atividade	Período
1o Semestre	02/03/2018 - 07/07/2018
Férias	08/07/2018 - 02/08/2018
2o Semestre	03/08/2018 - 08/12/2018
3o Semestre	01/03/2018 - 24/06/2018
Data Science Day	29/06/2018

Informações complementares

- ▶ moddle (moodle.c3sl.ufpr.br).
- ▶ Secretário: Sr. Valter.
- ▶ Coordenador: Wagner Hugo Bonat (wbonat@ufpr.br), Sala 227, LEG.
- ▶ Uso dos computadores: Prof. Daniel.
- ▶ Uso do clusters (computação de alto desempenho): Prof. Marco.