MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
CENTRO FEDERAL DE EDUCAÇÃO TECNOLÓGICA CELSO SUCKOW DA FONSECA
CONSELHO DE ENSINO, PESQUISA E EXTENSÃO

RESOLUÇÃO Nº 13/ 2014 **EM 06 DE NOVEMBRO DE 2014**

Aprova a Proposta de Criação do Programa de Pós-Graduação em Ciência de Dados - PPGCD

O Presidente do Conselho de Ensino, Pesquisa e Extensão do Centro Federal de Educação Tecnológica Celso Suckow da Fonseca, no uso de suas atribuições e em obediência à deliberação do CEPE, em sua 5ª. Sessão Ordinária, realizada em 06 de Novembro de 2014,

R E S O L V E:

Art. 1º - Aprovar a Proposta de Criação do Programa de Pós-Graduação em Ciência de Dados – PPGCD, conforme anexo.

Art. 2º - Esta Resolução entra em vigor na data de sua assinatura.

Carlos Henrique Figueiredo Alves
Presidente do Conselho de Ensino, Pesquisa e Extensão

Sumário

Programa de Pós-Graduação em Ciência de Dados

**Eduardo Ogasawara, Eduardo Bezerra, Jorge Soares,
João Quadros, Uéverton Souza**

Comissão de Criação do Programa
Grupo de Pesquisa em Computação Aplicada
EIC - Escola de Informática & Computação
Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – CEFET/RJ[1]

Resumo. *A Escola de Informática e Computação (EIC) - entidade conceitual formada pela Curso Técnico de Informática e pelo Departamento de Informática - apoiada pela Diretoria de Pesquisa e Pós-Graduação do CEFET/RJ, criou, em janeiro de 2011, o Grupo de Pesquisa em Computação Aplicada (GPCA). Esse grupo foi criado com o objetivo de fomentar a pesquisa na EIC, fortalecer o processo de criação do Bacharelado em Ciência da Computação e elaborar o processo de criação do Programa de Pós-Graduação em Ciência de Dados (PPGCD). Após quase quatro anos de dedicação e foco nessas atividades, julgamos ter corpo docente e estrutura suficientes para iniciar o processo de criação do PPGCD.*

1 Introdução

O mundo atual vive a era da Sociedade da Informação e do conhecimento, na qual empresas e centros de pesquisa compostos por pessoas capazes de agir com base na percepção e na relação de fatos globais assumem papel de relevância. Valoriza-se o capital intelectual, ativo dessas instituições, nem sempre concretamente materializado, mas que envolve o conhecimento sobre como realizar processos e tomar boas decisões nos diversos níveis institucionais.

Neste cenário, o ensino de Computação assume um papel de grande importância social, devendo formar profissionais que, além de uma boa base técnico-científica, possuam a capacidade de refletir, analisar, discernir e influir sobre as mais diversas questões do mundo contemporâneo, em particular àquelas relacionadas com as implicações da tecnologia computacional na sociedade. Afinal, a Informática se tomou uma realidade concreta e irreversível, cujo estágio tecnológico impõe uma presença que já não pode ser ignorada pela sociedade. A formulação de modelos computacionais que explicitem, incorporem e processem conhecimento também é uma característica desejável ao profissional de Computação.

A Computação está presente nos principais avanços em todas as áreas do conhecimento. Novas formas de interação entre as ciências, em vários níveis e escalas, são mediadas pela Tecnologia da Informação, que é a simbiose da Ciência da Computação com diferentes domínios do conhecimento. De fato, muitas das grandes descobertas científicas recentes são resultados do trabalho de equipes multidisciplinares que envolvem cientistas da Computação. A computação permeia todas as outras áreas

---

[1] Proposta a ser submetida à CAPES em maio de 2015. Início do curso previsto para março de 2016.

nas suas várias formas de investigação científica, tais como, simulação, modelagem, monitoramento, mensuração. Pode-se dizer que a Computação revolucionou a pesquisa científica, sendo hoje reconhecida como o "terceiro pilar" a sustentar a pesquisa, junto com os pilares da teoria e da experimentação.

De acordo com os dados apresentados no documento de área da CAPES de 2013 em Computação (CAPES 2013), o Brasil é o quarto maior mercado mundial de tecnologia da informação e comunicação (TIC) e sétimo maior em tecnologia da informação (TI). A expectativa é o país alcance a terceira posição em 2022. De acordo com a Associação Brasileira das Empresas de Tecnologia da Informação e Comunicação (Brasscom), estima-se que o setor de TIC tenha movimentado US$ 233 bilhões em 2012 e que alcance aproximadamente US$ 430 bilhões em 10 anos. O setor emprega hoje 2,5 milhões de pessoas. Nos próximos dez anos, a previsão é que venha a demandar mais um milhão de profissionais. Segundo o MCTI, apenas o mercado brasileiro de software deve crescer 400% nos próximos dez anos.

Esse cenário positivo gera demanda para formação de recursos humanos qualificados, o que exige planejamento e maiores investimentos. Além disso, para que o país alcance posições cada vez maiores em destaque internacional, é necessário um alto grau de inovação e pesquisa. Nesse sentido, é de interesse nacional a intensificação dos programas de pós-graduação que visem a formar mestres e doutores em Computação. Em particular há um demanda por profissionais cada vez mais capacitados em extrair conhecimentos a partir de grandes volumes de dados, os denominados cientistas de dados (DSC 2014).

Atualmente, pode-se dizer que diversas empresas urgem em contratar os cientistas de dados. Elas estão imersas no *dilúvio de dados* (do Inglês, *Data Deluge*) (Berman 2008), onde há grande volume de dados, com diferentes tipos de informação em uma escala sem precedentes. Essa demanda por estes especialistas de ciência da computação está bem à frente da capacidade de oferta. O tratamento do dilúvio de dados sendo produzido pelas ciências e por bilhões de usuários de serviços de Internet globais se apresenta como um dos grandes desafios para a atual sociedade do conhecimento (Bell et al. 2009). No mundo empresarial, os cientistas de dados são peças fundamentais para abordar o cenário de *Big Data* (Jagadish et al. 2014). Eles são capacitados a estruturar esses dados e encontrar padrões de modo a aconselhar os executivos sobre as implicações para produtos, processos e decisões (Dhar 2013).

No entanto, a demanda pelos cientistas de dados é bem mais ampla. O dilúvio de dados apresenta-se, de forma geral, em múltiplas facetas, fato que vem impulsionando iniciativas em diversas áreas, além do mundo empresarial, no sentido de mais bem entendê-lo. Nas ciências, o dilúvio apareceu como a expressão de uma nova maneira de investigação (Wright 2014), incentivando biólogos, astrônomos, físicos, e demais pesquisadores das mais diferentes áreas científicas a enfrentarem problemas computacionais na denominada e-ciência, que se tornam barreiras para as suas descobertas. Na setor governamental, há oportunidades de se debruçar sobre imensas bases de dados do setor público com vistas a gerar planejamento mais eficiente bem como novos serviços que possam melhorar o atendimento ao cidadão. O cientista de dados, é portanto, um profissional capacitado, principalmente, na análise, interpretação e manipulação de grandes volumes de dados, de modo a trazer o método científico para os mais diferentes setores.

Neste diapasão, o programa ora proposto é voltado ao mestrado *strictu-senso* e é intitulado *Programa de Pós-Graduação em Ciência de Dados (PPGCD)*. O viés aplicado e ao mesmo tempo científico do programa é uma característica presente no perfil profissional dos pesquisadores do quadro docente inicial proposto e esperado para os futuros postulantes a participar do programa. De fato, trata-se de uma característica presente na essência de nossa instituição: vincular a pesquisa a fins práticos e aplicados. Esta estratégia é ademais promissora, uma vez que ao mesmo tempo em que se estabelece resultados teóricos que subsidiam a construção de novas aplicações para solução de questões práticas, os problemas práticos muitas vezes propiciam a elaboração de novos arcabouços teóricos. Essa abordagem, baseada em Ciência Aplicada, adotada por nosso grupo está aderente ao processo multidisciplinar de aplicação da Computação às áreas estratégicas. As linhas de pesquisa propostas para o PPGCD são (i) Gerência de Dados e Aplicações; (ii) Métodos Baseados em Dados. Estas linhas e seus respectivos projetos são detalhadas ao longo deste documento.

Além desta introdução, este documento está organizado em mais treze seções. Nas seções 2 e 3, são apresentados, respectivamente, o histórico da instituição e a contextualização do PPGCD. Na seção 4 é apresentado o quadro inicial de pesquisadores do PPGCD. As seções 5, 6 e 7 apresentam, respectivamente, os objetivos, o perfil do egresso e a necessidades e a importância do curso. A estrutura curricular é abordada na seção 8, enquanto que a infraestrutura é abordada na seção 9. Nas seções 10, 11 e 12 são apresentadas, respectivamente, a política de integração com a graduação, intercâmbios e inserção social. A seção 13 apresenta a política de avaliação. Finalmente, a seção 14 conclui apresentando a política de credenciamento e a produção acadêmica do PPGCD.

2 Histórico

O CEFET/RJ – Centro Federal de Educação Tecnológica Celso Suckow da Fonseca – é uma Instituição Federal de Ensino Superior – IES – que tem por finalidade a oferta de Educação Tecnológica, tendo como objetivos: ministrar ensino em grau superior de graduação e pós-graduação *lato sensu* e *stricto sensu*; ministrar cursos técnicos, em nível de ensino médio, visando à formação de técnicos, instrutores e auxiliares; ministrar cursos de educação continuada visando à atualização e ao aperfeiçoamento de profissionais na área tecnológica; e realizar pesquisas na área tecnológica, estimulando atividades inventivas e estendendo seus benefícios à comunidade mediante cursos e serviços.

O CEFET/RJ incorpora o propósito de proporcionar a integração vertical entre os vários níveis de formação (médio/técnico, graduação e pós-graduação) e responsabiliza-se, ainda, pela qualificação docente para o ensino tecnológico no país, participando também da Universidade Aberta do Brasil – UAB.

Nos últimos anos, a Direção Geral do CEFET/RJ vem investindo fortemente na pesquisa e na formação de pesquisadores, estando ciente do papel estratégico do exercício de tais atividades dentro de um modelo universitário. O apoio à pesquisa e à pós-graduação pode ser observado por meio de ações como: criação, em 2007, da Diretoria de Pesquisa e Pós-Graduação – DIPPG (equivalente à Pró-Reitoria de Pesquisa e Pós-Graduação na estrutura de uma universidade); atualização/elaboração de regulamentação para pesquisa e pós-graduação na Instituição; e aumento significativo

da alocação de recursos próprios no centro de custos da DIPPG destinados à criação de infraestrutura adequada para atender às necessidades dos grupos de pesquisa e dos programas de pós-graduação. Outras ações importantes foram a adoção de critérios para alocação de recursos baseados em indicadores e o credenciamento anual de docentes para atuarem nos programas de pós-graduação.

O forte crescimento das atividades de pesquisa e pós-graduação no CEFET/RJ observado nos últimos anos pode ser traduzido pelo aumento expressivo da produção científica qualificada, do número de grupos de pesquisa, do número de programas de pós-graduação, do número de bolsistas de produtividade do CNPq, do número de bolsas de iniciação científica e de mestrado, além da ampliação da sua infraestrutura de pesquisa com a criação de novos laboratórios e a modernização dos existentes. A renovação do quadro docente nos últimos anos foi um fator essencial ao promover o aumento do corpo docente na Instituição, especialmente aqueles com titulação de doutor. Este panorama influencia diretamente nas perspectivas de evolução e consolidação do PPGCD.

Atualmente, a Instituição possui sete Programas de Pós-Graduação *Stricto Sensu*, oferecendo cursos de mestrado acadêmico, sendo que dois destes oferecem cursos de doutorado. Para maiores informações, veja http://dippg.cefet-rj.br/.

3 Contextualização do PPGCD

Em 2011, a Escola de Informática e Computação (EIC) - designação conceitual formada pelo Departamento de Informática (responsável pelos cursos superiores de Bacharelado em Ciência da Computação e de Tecnologia em Sistemas para Internet) e a Coordenação de Informática (responsável pelo Curso Técnico de Informática) - se organizou no sentido de estabelecer um planejamento estratégico para fomentar o ensino, a pesquisa e a extensão relacionados à Computação no CEFET/RJ. Naquela época, foi criado o Grupo de Pesquisa em Computação Aplicada (GPCA)[2]. O GPCA teve como missão impulsionar a pesquisa na EIC de modo a fortalecer o processo de criação do Curso de Bacharelado em Ciência de Computação (BCC), ocorrido em agosto de 2012, e direcionar os esforços no sentido de se criar o PPGCD.

Existe uma série de fatores que contribuíram para o fortalecimento do GPCA, a começar pela verba de pesquisa que o CEFET/RJ oferece aos grupos de pesquisas. Essa verba possibilitou a compra de equipamentos e a melhoria de infraestrutura laboratorial (ver seção 8). Tem havido também uma maior integração entre os membros do GPCA com outros pesquisadores das demais áreas (aplicadas) dentro da instituição (ver seção 10). O resultado disso pode ser observado inclusive na integração positiva com o ensino, onde há maior interesse por parte dos alunos do ensino médio-técnico e da graduação na participação em projetos de pesquisa (ver seção 9).

Com o GPCA criado e atuante desde 2011, pode-se observar um crescimento do grupo na área de pesquisa o que culminou na proposta do PPGCD. As linhas de pesquisa do PPGCD contemplam um amplo espectro de desafios em Ciência de Dados. A definição da área de concentração, bem como das linhas de pesquisas, foram refinadas ao longo do tempo. Em função da perspectiva ampla da Ciência de Dados

---

[2] http://dgp.cnpq.br/dgp/espelhogrupo/9806930220192669

(descrita na seção 3.1), a área de concentração é a Ciência da Computação. Dentro dessa área, foram organizadas duas linhas de pesquisa: (i) Gerência de Dados e Aplicações (seção 3.2); (ii) Métodos Baseados em Dados (seção 3.3). Cada uma dessas linhas é apresentada a seguir juntamente com respectivos projetos de pesquisa.

3.1 Ciência de Dados

Existe um grande debate no meio acadêmico e na indústria sobre a definição precisa do termo Ciência de Dados. Pode-se interpretar Ciência de Dados como uma evolução de áreas interdisciplinares que incorporam Ciência da Computação, Modelagem, Estatística de modo a extrair conhecimento a partir do processamento de grandes volumes de dados (NYU 2014). Neste contexto, constitui-se um desafio técnico-científico em computação o estudo metódico para a extração generalizada e em escala de conhecimento relevante a partir de uma imensa massa de dados, em geral dinâmicos (Jagadish et al. 2014). A abordagem a esse desafio com aplicações em diversas áreas no eixo ciência-indústria-governo emerge como uma nova espécie de ciência. A chamada Ciência de Dados incorpora elementos variados e se baseia em técnicas e teorias oriundas de muitos campos básicos em engenharia e ciências básicas, sendo assim intimamente ligada com muitas das disciplinas tradicionais bem estabelecidas, porém viabilizando uma nova área altamente interdisciplinar. Dessa forma, associado a este espírito de aplicação interdisciplinar, a ciência de dados se apresenta como componente cada vez mais importante nas mais diversas áreas relacionada, como saúde, petróleo, energia, finanças, esporte, astronomia, bioinformática, Internet, mobilidade urbana, defesa cibernética, comunicação móvel e biodiversidade, apenas para mencionar algumas.

Em ambiente altamente interdisciplinar com aplicações em áreas tão distintas, emerge o grande desafio comum às aplicações nessas tão diversas áreas de se identificar os princípios, métodos e técnicas fundamentais para o gerenciamento e análise de grandes volumes de dados, suplantando as dificuldades inerentes ao grande volume de dados em análise (Jacobs 2009, Lazer et al. 2014). Especificamente, identificamos duas linhas de pesquisa principais cujo amadurecimento julgamos conduzir rumo à consolidação da área de ciência de dados em um horizonte de alguns anos de pesquisa e desenvolvimento: (i) gerência de dados e aplicações; (ii) métodos baseados em dados. Todas essas linhas considerando a larga-escala dos dados a serem analisados bem como seu dinamismo.

Em uma visão geral, pode-se compreender a Ciência de Dados como um conjunto de ações aplicadas a uma coleção de dados que conduz à descoberta de conhecimento (i.e., tendências, relações, padrões subjacentes a esses dados). Dá-se o nome de *processo de ciência de dados* ao encadeamento de um conjunto de etapas, indicado na Figura 1 que começa com a seleção dos dados até a extração de conhecimento. As etapas que compõem o processo podem ser organizadas em quatro momentos: seleção, pré-processamento, análise e avaliação (Han and Kamber 2011). O processo pode ser compreendido como um caso particular de experimentação científica *in-silico* (Stevens et al. 2007) ou e-Science, no qual os dados são volumosos, as estruturas de dados precisam ser bem definidas e os métodos de seleção, pré-processamento e de análise de dados são computacionalmente intensivos.

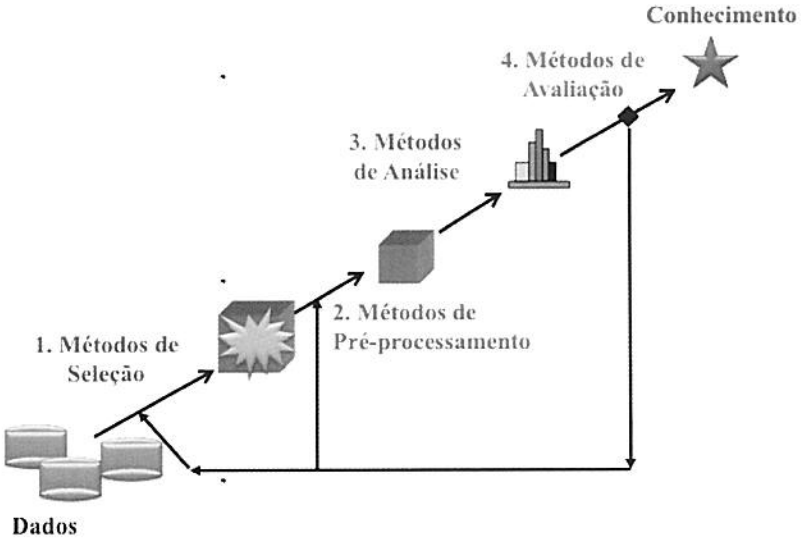

Figura 1 – Processo Iterativo de Ciência de Dados

A partir da Figura 1, pode-se observar as duas linhas de pesquisas. A linha de métodos baseados em dados compreendem todas as etapas de seleção de dados até a extração do conhecimento. A linha de gerência de dados e aplicações compreende todo o arcabouço que estabelece o experimento in-silico a partir de uma perspectiva datacêntrica.

Pode-se também observar as duas linhas pelas perspectivas de pesquisa básica e aplicada. A linha *métodos baseados em dados* é agnóstica ao domínio do problema, enquanto que a linha *gerência de dados e aplicações* é dependente do problema abordado e fortemente multidisciplinar. Destarte, os métodos baseados em dados têm um aspecto amplo no sentido de servir como pesquisa básica. A partir da pesquisa básica nesses aspectos fundamentais de gerência de dados e aplicações em larga escala, há também um grande potencial tecnológico na pesquisa aplicada em ciência de dados com impacto em diferentes áreas do conhecimento e em setores de atuação ao longo do eixo ciência-indústria-governo.

Um desafio correlato se torna a formação de recursos humanos altamente qualificados no desenvolvimento de pesquisa básica e aplicada na fronteira do conhecimento em ciência de dados. Esse cientista de dados possui demanda crescente no eixo ciência-indústria-governo (Davenport and Patil 2012). Esse profissional tem uma expectativa de formação tipicamente sólida em ciência da computação e aplicações, modelagem, estatística, analítica e matemática, além do conhecimento mínimo do domínio de aplicação.

Em suma, enfrentar de forma fundamental o grande desafio da ciência de dados permite contribuir de modo a melhor posicionar o Brasil na direção da nova ciência baseada em dados, preparando recursos hùmanos altamente qualificados, e desenvolvendo o alicerce para sua projeção de forma relevante na sociedade do conhecimento.

3.2 Gerência de Dados e Aplicações

Em linhas gerais, a Gerência de Dados e Aplicações corresponde ao arcabouço que descreve as etapas do experimento in-silico que devem ser desempenhadas, desde a

seleção dos dados até a produção de informação e do conhecimento. Nessa linha é apropriado estabelecer um tratamento datacêntrico a esses experimentos. A pesquisa nessa área envolve estruturas de dados e algoritmos diversos para apoiar o processo como um todo.

No que tange ao processo de ciência de dados, há também a necessidade premente de utilizar processamento de alto desempenho (PAD) para conseguir realizar a análise de dados em larga escala. Há importantes desafios no estabelecimento desses processos, comumente modelados como workflows (Deelman et al. 2009). Nesses workflows, as atividades e os dados estão direcionados à execução em algum ambiente de PAD (e.g., clusters, nuvens) (Ogasawara et al. 2013, de Oliveira et al. 2010). Em função da diversidade de plataformas existentes para ambientes de PAD, um dos grandes desafios é estabelecer uma representação desses workflows que seja agnóstica ao meio em que serão executados e, ao menos tempo, possibilite a otimização de sua execução no ambiente alvo.

Outro aspecto fundamental consiste em ter uma infraestrutura que possibilite explorar as diferentes técnicas para selecionar a mais adequada para os dados trabalhados. No processo de condução da Gerência de Dados e Aplicações, isso consiste em visualizar os resultados parciais e poder ajustar parâmetros nas técnicas de análise durante a execução do experimento (Mattoso et al. 2013).

Há inúmeros desafios nesta área relativos à elaboração de técnicas para análise de grandes volumes de dados (Berriman et al. 2007). As diferentes aplicações que demandam Gerência de Dados e Aplicações, ao mesmo tempo em que são alvos para elaboração de novas soluções em pesquisa aplicada, muitas vezes propiciam a oportunidade de elaboração de novos arcabouços teóricos em pesquisa básica, de caráter mais geral, para a solução dos problemas práticos. Essa abordagem, que liga teoria e prática, é uma das estratégias gerais adotadas pelos grupos pesquisa cujos estudos são centrados na análise de dados. Essa mesma abordagem será adotada pelo PPGCD.

Os principais projetos de pesquisa da linha Gerência de Dados e Aplicações do PPGCD compreendem *(i) Modelagem e Execução de Workflows para Ciência de Dados, (ii) Processamento de Dataflows sobre Grandes Volumes de Dados, (iii) Seleção de alvos para descoberta de fármacos, (iv) Mineração de dados de séries temporais, (v) Mineração de Dados para Sistemas de Detecção de Intrusão Baseados em Fluxos de Acesso, (vi) Avaliação de Artefatos de Aprendizado Baseados em Inteligência Computacional, (vii) Criação, Aplicação e Validação de Objetos de Aprendizagem Pervasivos.*

3.3 Métodos Baseados em Dados

Os métodos baseados em dados envolvem o processamento de coleções de objetos em busca de padrões consistentes na forma de relacionamentos sistemáticos entre variáveis componentes desses objetos, com o propósito de detectar e gerar conhecimento não facilmente detectado.

A extração do conhecimento propriamente dita é apoiada por um conjunto de métodos que incluem tanto os usados nas etapas tradicionais de mineração de dados - pré-processamento, identificação de padrões e visualização - quanto os métodos de reconhecimento de padrões e de modelagem computacional (Liao et al. 2012). Dentre

algumas dessas tarefas, podemos citar: agrupamento (*clustering*), indução de árvores de decisão, classificação e descoberta de associações.

A partir dessas tarefas clássicas, diversos algoritmos, implementações, adaptações e variações estão presentes em ferramentas de análise de dados consolidadas, como, por exemplo, a linguagem R. O desafio consiste tanto na correta aplicação desses algoritmos, como também na implementação adequada para se atingir escalabilidade no processamento de grandes volumes de dados. Esses componentes devem ser encapsulados em *wrappers* para que possam fazer parte dos experimentos de análise (workflows).

Nesse contexto, um aspecto muito importante de ligação com as técnicas de análise de dados consiste em como preparar os dados para a aplicação dessas técnicas. A correta aplicação das técnicas de normalização, transformação [Ogasawara et al. 2010], remoção de outliers [Gupta et al. 2014], seleção de atributos e definição de amostras, pode significar a diferença entre obter ou não conhecimento e produzir valor agregado.

Os principais projetos de pesquisa da linha Métodos Baseados em Dados compreendem *(i) Técnicas de Pré-processamento para mineração de dados, (ii) Teoria dos Grafos e Aplicações, (iii) Kernelização: Pré-processamento e Redução de Dados com Garantia, (iv) Simulação Numérica de Equações Diferenciais, (v) Métodos de imputação e dados* e *(vi) Métodos para extração de padrões em dados complexos.*

4 Quadro de Pesquisadores do PPGCD

O quadro inicial de pesquisadores do PPGCD é composto por doze docentes. Esse quadro tem, em sua maioria, professores da área de Ciência da Computação e pertencentes à unidade sede (Maracanã). Ele também possui professores de Computação das unidades de Nova Iguaçu e de Petrópolis. Dois dos docentes são de outras instituições Federais do Rio de Janeiro: Instituto Nacional de Metrologia, Qualidade e Tecnologia (INMETRO) e Laboratório Nacional de Computação Científica (LNCC).

Dos doze docentes, onze fazem parte do GPCA. De fato, o GPCA foi projetado para propiciar a futura criação do PPGCD. Nos próximos dois anos, seis docentes do GPCA devem obter o título de doutor e são potenciais candidatos a compor o quadro de pesquisadores do PPGCD. Além disto, outros professores das demais unidades também poderão se credenciar em breve. A seguir são apresentados os pesquisadores do quadro inicial do PPGCD.

Docentes Permanentes:

1. **Carlos Otávio Schocair Mendes (CEFET/RJ - Maracanã)**
   CPF: 856.863.287-49 | CV Lattes: http://lattes.cnpq.br/0093637819791265
   Titulação: Doutorado em Engenharia Elétrica (COPPE/UFRJ, 2010)
   Linhas de Pesquisa: Gerência de Dados e Aplicações
2. **Diego Nunes Brandão (CEFET/RJ – Nova Iguaçu)**
   CPF: 096.083.947-08 | CV Lattes: http://lattes.cnpq.br/5882024148867913
   Titulação: Doutorado em Computação (UFF, 2013)
   Linhas de Pesquisa: Métodos Baseados em Dados

3. **Eduardo Bezerra da Silva (CEFET/RJ - Maracanã)**
   CPF: 028.488.847-89| CV Lattes: http://lattes.cnpq.br/7568520840965379
   Titulação: Doutorado em Engenharia de Sistemas e Computação (COPPE/UFRJ, 2005)
   Linhas de Pesquisa: Métodos Baseados em Dados
4. **Eduardo Soares Ogasawara (CEFET/RJ - Maracanã)**
   CPF: 037.412.527-94 | CV Lattes: http://lattes.cnpq.br/0528303491410251
   Titulação: Doutorado em Engenharia de Sistemas e Computação (COPPE/UFRJ, 2011)
   Linhas de Pesquisa: Gerência de Dados e Aplicações
5. **João Roberto de Toledo Quadros (CEFET/RJ - Maracanã)**
   CPF: 796.161.787-68| CV Lattes: http://lattes.cnpq.br/0198821395183352
   Titulação: Doutorado em Ciência dos Materiais (IME, 2008)
   Linhas de Pesquisa: Gerência de Dados e Aplicações
6. **Jorge de Abreu Soares (CEFET/RJ - Maracanã)**
   CPF: 013.967.747-00 | CV Lattes: http://lattes.cnpq.br/3410221270317818
   Titulação: Doutorado em Engenharia de Sistemas e Computação (COPPE/UFRJ, 2007)
   Linhas de Pesquisa: Gerência de Dados e Aplicações
7. **Kele Teixeira Belloze (CEFET/RJ - Petrópolis)**
   CPF: 041.203.626-63| CV Lattes: http://lattes.cnpq.br/0309990426692102
   Titulação: Doutorado em Biologia Computacional e Sistemas (FIOCRUZ, 2013)
   Linhas de Pesquisa: Gerência de Dados e Aplicações
8. **Leonardo Silva de Lima (CEFET/RJ - Maracanã) – PQ2**
   CPF: 070.755.167-60 | CV Lattes: http://lattes.cnpq.br/0206233750299857
   Titulação: Doutorado em Engenharia de Produção (COPPE/UFRJ, 2006)
   Linhas de Pesquisa: Métodos Baseados em Dados
9. **Raphael Carlos Santos Machado (INMETRO) – PQ2**
   CPF: 093.691.287-19| CV Lattes: http://lattes.cnpq.br/9594450995231533
   Titulação: Doutorado em Engenharia de Sistemas e Computação (COPPE/UFRJ, 2010)
   Linhas de Pesquisa: Métodos Baseados em Dados
10. **Sérgio Eduardo Silva Duarte (CEFET/RJ - Maracanã)**
    CPF: 967.081.907-53 | CV Lattes: http://lattes.cnpq.br/8609836446570347
    Titulação: Doutorado em Física (CBPF, 2003)
    Linhas de Pesquisa: Gerência de Dados e Aplicações
11. **Uéverton dos Santos Souza (CEFET/RJ - Maracanã)**
    CPF: 057.816.147-88 | CV Lattes: http://lattes.cnpq.br/1131927171712708
    Titulação: Doutorado em Computação (UFF, 2014)
    Linhas de Pesquisa: Métodos Baseados em Dados

Docentes Colaboradores:

12. **Fabio André Machado Porto (LNCC) – PQ2**
    CPF: 884.045.957-04| CV Lattes: http://lattes.cnpq.br/6418711808050575
    Titulação: Doutorado em Informática (PUC-Rio, 2001)
    Linhas de Pesquisa: Gerência de Dados e Aplicações

5 Objetivos

O PPGCD tem como objetivos: (a) realizar pesquisa e desenvolvimento de métodos computacionais para a resolução de problemas complexos do mundo real; (b) formar pesquisadores altamente qualificados na área da Ciência de Dados que possam atuar tanto no mercado quanto na academia; (c) estimular o intercâmbio de conhecimento e a inovação tecnológica entre a academia e a sociedade.

6 Perfil do Egresso

O planejamento do perfil do do PPGCD prevê grande atuação nos setores onde há necessidade de extração de conhecimento a partir dos dados. Em especial nas áreas de educação, serviços, instituições financeiras, petróleo e gás, defesa e segurança cibernética, áreas estratégicas, tanto pela demanda própria do Estado do Rio de Janeiro, quanto pela percepção no documento de área da Computação.

Para tanto, é necessário que o curso em tela prepare seus discentes com substancial formação teórico-prática na linha de pesquisa em que estão vinculados. De maneira geral, planejamos que o egresso desse curso detenha conhecimento que o habilite tanto para tratar de questões teórico-práticas relacionadas aos problemas mais atuais relacionados à Computação, quanto tenha habilidade de propaga-las no meio acadêmico-científico, seja em cursos de qualificação, aperfeiçoamento, ou mesmo na formação clássica em nível de graduação e pós-graduação.

7 Necessidade e Importância

O PPGCD visa à formação de recursos humanos qualificados na área de Ciência da Computação enfatizando a Ciência de Dados. Em particular, visa à formação de pesquisadores capazes de resolver de problemas do mundo real que estabelecem um ciclo virtuoso entre as pesquisas aplicada e básica. Essa abordagem apresenta forte iteração com a sociedade. Neste cenário, o PPGCD será o primeiro programa de pós-graduação do Brasil em Ciência de Dados, o que abre um espaço de liderança do CEFET/RJ no cenário brasileiro. Pode-se indicar a necessidade e importância do PPGCD pelas dimensões institucionais, de contextualização e de demanda.

7.1 Dimensão Institucional

Em agosto de 2012 o curso de Bacharelado de Ciência da Computação (BCC) foi criado na sede de nossa instituição (Maracanã). A partir desta data, estabeleceu-se o planejamento de criação do mestrado para iniciar a operação a partir de março de 2016, período no qual a primeira turma do BCC se formará. Junte-se a esse fato a crescente demanda interna de nossa instituição, que hoje conta com diversos cursos relacionados à Computação: Curso Superior de Tecnologia em Sistemas para Internet (unidade Maracanã), Engenharia da Computação (unidade Petrópolis) e Sistemas de Informação (unidade Friburgo).

A formação do PPGCD com quadro docente proveniente dessas diferentes unidades fortalece a instituição como um todo, atende aos alunos que pretendem continuar seus estudos e também fortalece os projetos pedagógicos das graduações relacionadas, uma vez que há um potencial de forte interação entre os níveis de graduação e de pós-graduação.

Cabe ressaltar também a presença de outros programas do CEFET/RJ, como o Programa de Pós-Graduação em Tecnologia (PPTEC), Programa de Engenharia Elétrica (PPEEL), Programa de Engenharia Mecânica e Tecnologia de Materiais (PPEMM) e Programa de Ensino de Ciências e Matemática (PPECM). Esses programas podem ter forte associação com o PPGCD, onde é forte a sinergia da computação como uma proposta de Ciência de Dados. De fato, esta situação já ocorre entre alguns pesquisadores do GPCA e do PPTEC.

7.2 Contextualização frente a outros Programas

No Estado do Rio de Janeiro, existem mais seis programas de pós-graduação em Ciência da Computação *strictu-senso* em instituições de ensino públicas federais, sendo dois na Universidade Federal do Rio de Janeiro (UFRJ), um na Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), um na Universidade Federal Fluminense (UFF), um na Universidade Federal do Estado do Rio de Janeiro (UNIRIO) e um no Instituto Militar de Engenharia (IME). Em função disso, apresentamos uma contextualização desses demais programas existentes e posicionamos o PPGCD frente a eles.

O Programa de Engenharia de Sistemas e Computação (PESC) da COPPE/UFRJ tem como visão ser reconhecido como o melhor programa de pós-graduação em Computação da América Latina, tendo seu nome reconhecido internacionalmente como um dos melhores do mundo. O PESC é um programa de excelência (nível 7 da CAPES) e com grande diversidade de áreas de concentração: (i) Algoritmos e Combinatória; (ii) Arquitetura e Sistemas Operacionais; (iii) Banco de Dados; (iv) Computação Gráfica; (v) Engenharia de Software; (vi) Engenharia de Software; (vii) Informática e Sociedade; (viii) Inteligência Artificial; (ix) Otimização; (x) Redes de Computadores.

O Programa de pós-graduação do Departamento de Informática (PPGDI) da PUC-Rio é também de excelência (nível 7 da CAPES). O PPGDI apresenta as seguintes linhas de pesquisa: (i) Banco de Dados; (ii) Computação Gráfica; (iii) Engenharia de Software; (iv) Linguagens de Programação; (v) Redes de Computadores e Sistemas Distribuídos; (vi) Teoria da Computação; e (vii) Otimização e Raciocínio Automático; (viii) Hipertexto e Multimídia; e (ix) Interação Humano-Computador. Dessas linhas, as sete primeiras podem ser consideradas clássicas, enquanto que as duas últimas têm perfil de inovação. Essa instituição tem uma sólida tradição na área de multimídia, possuindo linguagens e padrões mundialmente conhecidos na área. O PPGDI inclusive coordena e abriga o INCT de *Web Science*.

O Programa de Pós-Graduação em Computação (PGC) da UFF tem por objetivo principal a formação de profissionais altamente qualificados para as atividades de ensino e pesquisa em Computação, incluindo suas aplicações práticas, em particular na Engenharia. Atualmente é um programa classificado no nível 6 da CAPES e organizado em seis áreas de concentração: (i) Algoritmos e Otimização; (ii) Computação Científica e Sistemas de Potência; (iii) Computação Visual; (iv) Engenharia de Software; (v) Inteligência Artificial; e (vi) Redes e Sistemas Paralelos e Distribuídos. Dentro destas áreas, o PGC possui diversas linhas de pesquisas, algumas das quais destas são relacionadas ao PPGCD: Mineração de Dados, Teoria e Algoritmos em Grafos.

O Programa de Pós-Graduação em Informática (PPGI/UNIRIO) da UNIRIO tem como área de concentração a pesquisa em Sistemas de Informação, o que se apresenta como diferencial estratégico desse programa no contexto do Rio de Janeiro e do Brasil. O PPGI é um programa nível classificado no nível 4 da CAPES e tem agido como catalisador do reconhecimento de Sistemas de Informação como área de pesquisa no cenário nacional, acompanhando e gerando tendências e pesquisas de ponta na área.

O Programa de Pós-Graduação em Informática (PPGI/UFRJ) da UFRJ tem como filosofia básica expor os alunos a problemas derivados de aplicações em diferentes áreas do conhecimento que necessitem, para a sua solução, da superação de desafios concretos, que estimulem a criatividade, e gerem soluções inovadoras no contexto da Informática. O mestrado tem cinco linhas de pesquisa básicas: (i)

Algoritmos e Métodos Numéricos; (ii) Informática, Educação e Sociedade; (iii) Modelos e Arquiteturas para Sistemas Inteligentes; (iv) Redes de Computadores e Sistemas Distribuídos; e (v) Sistemas de Informação.

O Programa de Pós-Graduação em Sistemas e Computação (PGSC) do Instituto Militar de Engenharia (IME) tem como missão contribuir para a melhoria e aperfeiçoamento da área de Ciência de Computação do país, a partir das demandas do Exército Brasileiro (EB) e da sociedade. O PGSC tem uma única área de concentração: Ciência da Computação. As linhas de pesquisa do programa são: (i) Metodologias Computacionais; (ii) Sistemas Computacionais; (iii) Sistemas de Informação. A instituição está em processo de reformulação dessas linhas, e há a indicação do estabelecimento de um foco voltado à Estruturação da Pesquisa Científica na Área Cibernética do Exército Brasileiro.

PESC, PPGDI e PGC são tradicionais e de grande relevância nacional. Nas suas respectivas instituições, alguns dos pesquisadores e parte das linhas de pesquisas têm interseção com o PPGCD. Isso é natural, tendo em vista o amplo espectro desses programas. Entretanto, a proposta do PPGCD difere essencialmente pelo escopo e pela abordagem metodológica. As linhas de *Gerência de Dados e Aplicações* e *Métodos Baseados em Dados* não estão presentes nos três programas supracitados. A abordagem aplicada e focada nos domínios de aplicação é essência do PPGCD. Nossa filosofia é explorar as questões teóricas a partir das aplicações. Essa abordagem viabiliza uma maior sinergia entre as linhas de pesquisas do PPGCD como também visa à formação de um perfil diferenciado de aluno: o cientista de dados.

Em uma visão mais ampla, as linhas de pesquisa do PPGCD não possuem interseção com as existentes no PPGI/UNIRIO, no PGSC e no PPGI/UFRJ. Além disso, o PPGI/UNIRIO e o PGSC são programas especializados. O primeiro é focado em uma subárea do conhecimento da computação (Sistemas de Informação), enquanto que o segundo tem a sua estrutura de linhas voltadas a atender os desafios da Área Cibernética do Exército Brasileiro. O PPGI/UFRJ possui uma diversificação de atuação utilizando uma abordagem aplicada, o que é semelhante ao PPGCD, mas não tem interseções com as linhas de pesquisa e os objetivos do PPGCD.

A diferença para o PPGCD em relação aos demais programas de nosso Estado é clara. O PPGCD atua na formação de cientistas de dados e propõe se tornar o primeiro programa de pós-graduação do Brasil em Ciência de Dados.

7.3 Dimensionamento da Demanda

Em relação à demanda da área de conhecimento, o próprio documento de área indica a Computação como área estratégica e em crescente expansão. Os dados a seguir foram coletados a partir de um estudo realizado pela *Brasscom* (Associação Brasileira de Empresas de Tecnologia da Informação e Comunicação). Neste estudo, é também apresentada a expectativa de demanda por profissionais de TI no Rio de Janeiro, que se apresenta como a segunda maior do país, ficando atrás de São Paulo apenas. Além disso, a competividade mundial na área de TI gera demanda por profissionais cada vez mais capacitados em Computação, que são candidatos a realizarem pós-graduação. A Tabela 1 apresenta a formação de graduados em Computação e a demanda por profissionais na área.

Tabela 1 - Oferta e Demanda de profissionais de TI no Rio

	2007-2009	2010	2011	2012	2013	2014
Graduados	3474	4555	4600	4692	4833	5026
Demanda	2193	2912	3714	4840	6491	9076

Apesar de existirem outras instituições de pós-graduação no estado do Rio de Janeiro, a perspectiva de expansão da Computação no Brasil aumenta a demanda para criação de novos programas. Atualmente são formados anualmente por volta de 178 mestres provenientes de pós-graduações da área de Computação no Estado do Rio de Janeiro. Conforme a Tabela 2, nestes últimos anos, houve uma expansão do corpo docente da UFF. Da mesma forma, a UNIRIO apresentou uma proposta com um nicho bem caracterizado.

Um aspecto importante a observar é que a inserção da UNIRIO e da UFF como instituições de pós-graduação em Computação não impactou negativamente nas demais pós-graduações. Partimos desta premissa também para apoiar a nossa proposta. Nossa premissa é embasada nas demandas nacionais e estaduais de formação de profissionais de TI. Alia-se a esse estudo a própria demanda interna presente em nossa instituição.

Tabela 2 – Mestres formados nas IFES do Rio de Janeiro

IFES	2008	2009	2010	2011	2012
PESC (COPPE/UFRJ)	42	34	44	43	41
PPGDI (PUC-RIO)	40	48	37	44	32
PGC(UFF)	14	19	17	24	32
PGSC (IME)	19	12	9	9	16
PPGI (UNIRIO)		24	21	23	23
PPGI (UFRJ)	23	28	29	30	34
Total	138	165	157	173	178

Sumarizando, conforme indicado neste documento, o PPGCD estabelece linhas de pesquisa e abordagens que nos diferenciam de modo a contribuir e complementar as demais pós-graduações em Computação de nosso Estado e, ao mesmo tempo, atender a um conjunto diferenciado de alunos de Computação e a uma demanda crescente por profissionais de Ciência de Dados.

8 Estrutura Curricular

O curso de mestrado do PPGCD é composto de um elenco de disciplinas eletivas divididas em conformidade com as três linhas de pesquisa. Os estudantes devem integralizar, pelo menos, 24 (vinte e quatro) créditos distribuídos de tal modo que cumpram, no mínimo, 3 (três) créditos em cada linha de pesquisa. No caso de estudantes bolsistas, será obrigatório o cumprimento da disciplina *Estágio de Docência*.

Os discentes também precisam cursar as disciplinas "Seminário para Dissertação" e "Pesquisa para Dissertação", ambas sem atribuição de créditos, mas obrigatórias. Na disciplina Seminário para Dissertação os alunos apresentam a proposta de dissertação, que deve ser aprovada por uma Banca Examinadora.

O curso é organizado em regime trimestral. As disciplinas são apresentadas na Tabela 3.

Tabela 3 – Disciplinas do PPGCD

Disciplina	Créditos
Metodologia em Pesquisa	3
Métodos Estatísticos	3
Mineração de Dados	3
Aprendizado de Máquina	3
Gerência de Dados e Aplicações	3
Algoritmos e Estruturas de Dados	3
Tópicos em mineração de dados	3
Tópicos em aprendizado de máquina	3
Tópicos em gerência de dados e aplicações	3
Tópicos em algoritmos e estruturas de dados	3
Seminário para a Dissertação de Mestrado	0
Pesquisa para a Dissertação de Mestrado	0

9 Infraestrutura

9.1 Laboratórios

Os alunos do PPGCD irão dispor da seguinte estrutura de laboratórios: (i) cinco laboratórios de ensino já existentes na EIC; (ii) um laboratório de pesquisa em Ciência de Dados (LPCA), também já montado; (iii) um laboratório de uso geral. Cada um desses laboratórios tem aproximadamente 20 computadores.

Os laboratórios são multiuso com espaço para projeção multimídia e uso de quadro branco. Os computadores do laboratório possuem diversos ambientes de programação (Java, Python), clientes de Banco de Dados (PostgreSQL) e ferramentas como R, Libre Office, LaTeX. Os computadores presentes em quatro laboratórios tem o Sistema de Workflows Sagitarii, o que os torna um cluster Beoulful.

Há um conjunto de servidores PowerEdge que servem de *backend* (servidores de banco de dados e de aplicações) para as máquinas presentes nos laboratórios e 100% dedicadas às atividades de ensino e pesquisa. Há também um sistema de tele-conferência disponível e que se encontra instalado na sala de reuniões da Diretoria de Pesquisa e PósGraduação – DIPPG.

Esses laboratórios têm recebido, ao longo dos últimos anos, investimentos que permitiram a aquisição de servidor de alto desempenho e de novos equipamentos de multimídia, a atualização dos computadores e a troca do mobiliário e dos condicionadores de ar (*splits*). A atualização da infraestrutura dos laboratórios da EIC tem sido feita com recursos próprios do CEFET/RJ (seja por pedido direto da EIC, seja por pedido do GPCA) e com recursos advindos de projetos de pesquisa financiados por órgãos de fomento como a FAPERJ.

9.2 Bibliotecas

A Biblioteca Central do CEFET/RJ funciona no quarto andar do Bloco E. Além do espaço individual de leitura, conta com sala de estudos, dois miniauditórios, um auditório maior, um setor de multimídia, áudio e vídeo e um setor para consulta virtual.

Desde 2010, quando foi implantado o Sistema Phoenix de Bibliotecas, encontra-se disponibilizada a consulta online ao acervo das bibliotecas que compõem o sistema CEFET/RJ (Biblioteca Central e bibliotecas das Unidades Descentralizadas – UNEDs). Ao longo dos últimos anos, a Biblioteca vem contando com investimentos constantes para ampliação de seu acervo. No que se refere especificamente as demandas do PPGCD, os títulos poderão ser adquiridos através de pedidos da EIC e do GPCA. Há que se ressaltar que a verba destinada à pesquisa no CEFET/RJ é distribuída entre os diversos grupos de pesquisa mediante critérios estabelecidos em edital, possibilitando aos grupos a aquisição de bibliografia, equipamentos e serviços necessários para a realização das atividades de pesquisa. Há que se destacar o Portal de Periódicos da CAPES que tem suprido muitas das necessidades de pesquisa bibliográfica do curso, podendo ser acessado da Sala dos Alunos ou de qualquer outro computador da Instituição.

10 Integração com a graduação

O processo de criação do PPGCD emana de iniciativas provindas dos colegiados do Curso Técnico de Informática e das graduações em computação da EIC. Por essa razão, o corpo docente do PPGCD tem foco nas relações entre esses três níveis de ensino. Tal iniciativa é aderente ao PDI do CEFET/RJ. Especificamente no caso do PPGCD, todos os docentes da instituição, sem exceção, interagem com a graduação e/ou curso técnico ministrando aulas, orientando projetos de iniciação científica e projetos de fim de curso. Além da oferta de disciplinas obrigatórias constantes da grade curricular dos cursos de graduação, foram criadas disciplinas eletivas introdutórias a outras disciplinas previstas no mestrado, com o objetivo de estimular o interesse, preparar e aproximar os alunos de graduação das atividades desenvolvidas no PPGCD.

A oferta de disciplinas para a graduação também fica assegurada pelas normas institucionais de credenciamento e manutenção de credenciamento docente em programas de pós-graduação stricto sensu do CEFET/RJ, que estabelecem que todo docente, para atuar na pós-graduação precisa ministrar, no mínimo, uma disciplina na graduação ou ensino médio/técnico (quando for o caso).

Com relação à iniciação científica, o CEFET/RJ oferece dois programas, a Iniciação Científica (PIBIC) e o Iniciação Científica Ensino Médio-Técnico (PIBIC-EM): Anualmente é publicado um edital para concessão de bolsas de iniciação científica custeadas pelos Programas PIBIC e PIBIC-EM. A instituição possui 80 bolsas de iniciação científica (30 financiadas pelo CNPq e 50 financiadas pelo CEFET/RJ) e 40 bolsas de iniciação científicq ensino médio técnico (20 financiadas pelo CNPq e 20 financiadas pelo CEFET/RJ). Mesmo os alunos não contemplados com a concessão de bolsa podem realizar iniciação científica, desde que cumpram todas as atividades e exigências do programa (submissão de projeto no período estabelecido pelo edital, frequência, apresentação de relatórios e participação no Seminário de Iniciação Científica e Tecnológica). O PIBIC é acompanhado por um comitê interno e por um comitê externo, o qual é composto por pesquisadores do CNPq. Em 2006-07 (última avaliação dos programas PIBIC disponibilizada pelo CNPq), o PIBIC do CEFET/RJ foi classificado em 5º lugar na lista de todas as instituições brasileiras avaliadas. A distribuição das bolsas é feita com base nos critérios de classificação vigentes estabelecidos pelo comitê local. Os resultados dos projetos são apresentados pelos

alunos no Seminário de Iniciação Científica e Tecnológica do CEFET/RJ, evento anual promovido pela Instituição, e os resumos são publicados no Livro de Resumos editado pela gráfica da Instituição. Em 2013, o Seminário de Iniciação Científica e Tecnológica foi realizado na Semana de Extensão juntamente com o Seminário de Pesquisa e Pós-Graduação, no evento intitulado Seminário de Pesquisa e Pós-graduação 2013, que contou com o apoio da FAPERJ por meio do Programa APQ2 (Apoio a Eventos).

Finalmente, a partir da última reunião do Conselho de Ensino Pesquisa e Extensão (CEPE) do CEFET/RJ, realizada em 02 de outubro de 2014, foi aprovada a criação da revista Cadernos em Ciência de Dados (CCA). O corpo editorial da revista é composto por pesquisadores da UFF, UFRJ, LNCC, IME e INMETRO. O objetivo dessa revista eletrônica é fomentar a publicação de artigos originários de Trabalhos de Conclusão de Curso (TCC) e Iniciações Cientificas (IC). Esperamos que essa revista estimule os alunos no exercício da publicação científica e tecnológica, além de envolvê-los cada vez mais em atividades de pesquisa. Em particular, a publicação de artigos resultantes de seus TCCs deve incentiva-los a elaborar trabalhos de fim de curso mais qualificados e a continuar seus estudos em uma pós-graduação.

11 Intercâmbios

Desde 2011, os pesquisadores do PPGCD, por meio do GPCA, têm sistematicamente promovido ciclos de palestras no denominado Workshop em Ciência de Dados (WCA). Participam como palestrantes convidados no WCA professores renomados de nosso Estado. Esses ciclos têm o potencial de aproximar o PPGCD a pesquisadores e programas de pós-graduação de outras instituições. Esses ciclos têm sido enriquecedores para a formação dos alunos, tanto do nível técnico quanto do nível superior, fomentando o desejo pelo empreendedorismo, pela pesquisa e inovação e pela publicação de trabalhos científicos em veículos qualificados, além de promover um intercâmbio institucional.

No que tange a colaborações interinstitucionais, o GPCA tem parcerias com diversos centros de pesquisa, dos quais destacamos o Laboratório Nacional de Computação Científica (LNCC), onde fazemos parte da proposta de criação de INCT em Gerência de Dados e Aplicações. Pode-se destacar também a parceria com o Observatório Nacional por meio do LIneA (Laboratório Interinstitucional de e-Astronomia), onde também fazemos parte da proposta de criação do INCT do e-Universo. Temos também interações com outras instituições de pesquisa, como o Centro de Tecnologia Mineral (CETEM) e com outras IFES: Universidade Federal do Rio de Janeiro (UFRJ), Universidade Federal Fluminense (UFF), Universidade do Estado do Rio de Janeiro (UERJ) e Instituto Militar de Engenharia (IME).

No que tange à interação com a Indústria, pode-se destacar a interação de pesquisa com as empresas Clavis Segurança da Informação (detecção de invasão), EUD Tecnologia da Informação (engenharia de software) e Finxi Tecnologia (processamento de imagens). As empresas trazem problemas complexos que demandam a aplicação do método científico para sua resolução. Os alunos interessados em participar desses problemas podem desenvolver temas de Trabalho de Conclusão de Curso (TCC) ou de iniciação científica. Os resultados podem tanto se transformar em artigos científicos quanto serem absorvidos pela empresa associada.

Finalmente, na área de pesquisa em Informática na Educação, o CEFET/RJ está direcionando os esforços de pesquisa aplicada ao trabalho com escolas do município do Rio de Janeiro. Iniciou-se um projeto piloto junto ao Colégio Olímpico Juan Antônio Samaranch, no bairro de Santa Teresa, onde se pretende levantar demandas de apoio computacional e realizar inúmeros experimentos de modo a estimular os alunos a durante o processo de ensino-aprendizagem.

11.1 Internacionalização

Outros intercâmbios têm sido propiciados por meio de projetos de pesquisas, com apoio de órgãos de fomento, firmados entre professores do PPGCD e pesquisadores de outras instituições. No que tange a projetos internacionais, pode-se destacar o projeto MUSIC (gerência de dados científicos em uma nuvem multi-site) do Programa FAPERJ-INRIA, envolvendo cooperação entre LNCC e INRIA, França. O coordenador do lado brasileiro é o Prof. Fábio Porto (LNCC) e o coordenador do lado francês o pesquisador Patrick Valduriez (INRIA). Participa deste projeto representando o CEFET/RJ o Prof. Eduardo Ogasawara.

Outras ações de internacionalização:

- Organização o VI Workshop de e-Science no CSBC. Apesar de o evento ser nacional, o comitê de programa foi composto por diversos pesquisadores internacionais. Além disto, tivemos a participação de palestrantes internacionais de relevância, como a Prof. Anatasia Ailamaki da Ecole Polytechnique Fédérale de Lausanne e David Schade, líder do grupo do Canadian Astronomy Data Centre (CADC);
- Projetos relacionados ao estudo de problemas combinatórios e de Métodos em parcerias com os professores da Universidade de Ulm na Alemanha, mais especificamente com os pesquisadores Dieter Rautenbach e Lucia Draque Penso do Instituto de Otimização e Métodos (Universität Ulm - Institut für Optimierung und Operations Research). A parceria com o Instituto de Otimização e Métodos da Universidade de Ulm se estende desde o doutorado do prof. Uéverton Souza, o qual foi orientado pelo prof. Dieter Rautenbach na Alemanha durante seu doutorado sanduíche.

12 Inserção Social

A inserção social do PPGCD estará inicialmente voltada às atividades de ensino. A linha de Gerência de Dados e Aplicações terá um papel bastante importante nesta interação. Em particular, as aplicações voltadas ao ensino, serão avaliadas em instituições públicas de ensino fundamental e médio do município do Rio de Janeiro. Um exemplo disso é o projeto de desenvolvimento no GEO (Ginásio Experimental Olímpico) de Santa Teresa, onde estamos trabalhando em conjunto de ensino de robótica. Estas técnicas promoverão maior inclusão social e ao mesmo tempo poderão servir como motivadores a que estes estudantes considerem a Computação como potencial carreira para seguir os estudos no ensino médio-técnico ou na graduação.

Outra ação mais geral consiste na organização de eventos abertos à comunidade, como os eventos organizados nas *semanas de pesquisa e extensão*, onde são

apresentadas diversas palestras na área da computação, tanto de teor acadêmico quanto de teor empreendedor e mercadológico. Esse tipo de evento é um potencial atrator para área de computação.

13 Avaliação

Além das avaliações normais referentes a cada uma das disciplinas realizadas, no fim do curso, cada estudante deverá apresentar e defender tanto uma dissertação circunscrita em uma das linhas de pesquisa, quanto um material didático, paradidático ou processos educacionais com o seu respectivo planejamento teórico vinculado à dissertação.

14 Política de Credenciamento e Produção Acadêmica do PPGCD

14.1 Critério de Credenciamento

A área de Ciência da Computação entende como periódicos os veículos de divulgação com corpo editorial reconhecido, com avaliação pelos pares (pareceristas *ad hoc*), dotados de ISSN e que aparecem em bases de dados reconhecidas internacionalmente. As fontes de dados mais relevantes para a área são: ISI, Scopus, ACM, IEEE, SpringerLink, InterScience, ScienceDirect e Scielo.

A área também considera de igual importância a publicação de artigos completos em anais de conferências tradicionais quanto a publicação em periódicos. Essas conferências contam com comitês de programa e realizam um processo rigoroso de avaliação por pares. Cabe ressaltar que, de acordo com a política de documentos da Scopus, a definição de artigo engloba tanto as publicações constantes em periódicos quanto em anais de conferências.

Essas conferências, na sua maioria, constam das mesmas fontes citadas acima e, conforme indicado em estudo feito pela área de Ciência da Computação, podem ser avaliadas seguindo os mesmos índices e parâmetros dos periódicos. A área de computação também apresenta saturação (travas) em relação a publicações de conferência. O número total de artigos em conferências contabilizáveis para um programa está limitado a três vezes o número de artigos em periódicos para o período de avaliação. Essa limitação é coerente com o padrão de publicações de centros de excelência no exterior na área de Ciência da Computação, que publicam em média 2,5 artigos em conferências para cada artigo em periódico.

De acordo com o documento de área da Capes em Computação, a avaliação da produção intelectual é feita por meio de dois índices I_{geral} e $I_{restrito}$. O índice $I_{restrito}$ leva em consideração apenas os veículos de maior classificação no Qualis (A1 a B1). Este índice equivale a uma saturação para os estratos B2 a B5. A Tabela 4 apresenta a pontuação de pesos por Qualis.

Tabela 4 – Tabela de Pesos para Qualis

A1	A2	B1	B2	B3	B4	B5	C
100	85	70	50	20	10	5	0

À luz desse contexto, a política inicial do PPGCD para credenciamento dos docentes irá adotar os seguintes critérios: o pesquisador deve possuir 140 pontos, sendo necessariamente 70 pontos na faixa restrita (i.e., A1 até B1).

14.2Produção do PPGCD

De modo a posicionar a proposta do PPGCD por uma perspectiva quantitativa, foi feita uma avaliação da produção dos pesquisadores que formam o quadro inicial do programa proposto, levando em consideração o triênio de 2012-2014. A produção total de artigos (medida por docentes permanentes) foi de 113 artigos. Dado o quadro inicial de 11 docentes permanentes e o período de análise (3 anos), a produção média anual de cada docente é de 3,1 artigos. Esse é um resultado bastante positivo para a candidatura do PPGCD. A Tabela 5 apresenta a produção dos membros permanentes do quadro docente inicial do PPGCD, considerando-se os cenários geral (i.e., 10 docentes internos e 1 docente externo) e interno (i.e., apenas os 10 docentes internos) para o triênio 2012-2014.

Pode-se observar também que uma parcela significativa da produção em revistas está concentrada nos índices restritos. Em termos de números totais, a proporção entre artigos de revistas e congressos também está aderente ao que está descrito no documento de área. Uma consideração a ser mencionada está na taxa de artigos, tanto em revista (37%), mas principalmente em congressos (56%) que não apresentam Qualis. Este indicador foi observado e será trabalhado pelo grupo nos próximos anos. Nos dados apresentados a seguir, note que não é considerada a produção do docente colaborador.

Tabela 5 – Tabela Produção de Congressos e Revistas por Qualis

	A1	A2	B1	B2	B3	B4	B5	C	Total
Congressos (cenário geral)		1	8	6	2	13	2	40	**72**
Revistas (cenário geral)	4	11		5	2		4	15	**41**
Congressos (cenário interno)		1	5	5	2	11	2	29	**55**
Revistas (cenário interno)	1	6		2	2		3	14	**28**

A Tabela 6 apresenta a produção individualizada de cada docente, considerando-se toda a produção. Cabe ressaltar que, para os membros internos, esta produção está predominantemente concentrada nos últimos anos e vem crescendo.

Tabela 6 – Produção Acadêmica dos Pesquisadores

	Artigos Comp.	Anais Comp.	Anais Resumo	Capitulo	Livro	IC-Grad	TCC-Grad	Mestrado	Doutorado
Carlos Otávio Schocair Mendes	1	4	7				4		
Diego Nunes Brandão	4	16	4			4	16		
Eduardo Bezerra da Silva	3	16	1	3	1	11	32	1	
Eduardo Soares Ogasawara	17	65	4	1		22	7	2	
João Roberto de Toledo Quadros	4	17	5			2	3		
Jorge de Abreu Soares	3	14	3	2	3		23	2	
Kele Teixeira Belloze	1	4	7				5		
Leonardo Silva de Lima	10	18	5			6	14	9	1
Sérgio Eduardo Silva Duarte	8	3		2				7	
Uéverton dos Santos Souza	5	11	3			2	6		
Raphael Carlos Santos Machado (externo - permanente)	20	32	5	2				1	1
Fabio Andre Machado Porto (externo – colaborador)	17	48	6	2	3	2	4	20	2

14.3Comparativo com os programas da área de Ciência da Computação

Atualmente existem *68* programas de pós-graduação na área de Ciência da Computação em todo país, dentre os quais seis são pertencentes ao estado do Rio de Janeiro. Desses *68* programas, *35* são classificados como 3, *20* como 4, *5* como 5, *3* como 6, e *5* como 7 pela CAPES. Com o intuito de comparar a produção científica do quadro docente inicial do PPGCD frente a esses programas de pós-graduação, analisamos os índices de produções qualificadas dos programas por docente (PQD) e do grupo proposto para o quadro inicial do PPGCD.

Embora na área de computação trabalhos publicados em conferências também pontuem para os programas, devido à insuficiência de informações, em nossa análise consideramos apenas as publicações em periódicos. No entanto, do ponto de vista de avaliação junto à CAPES, o número de publicações em conferências é saturado por 3 vezes o número de publicações em periódicos. Portanto, a posição relativa dos programas apresentada tende a ser próxima do cenário completo, em que as informações sobre conferências seriam consideradas.

Para fins de análise, consideramos tanto o índice restrito *IR* (levando em conta apenas periódicos qualificados entre A1 e B1) quanto o índice irrestrito *II* (considerando periódicos qualificados entre A1 e B5).

$$IR = (A1 + A2*0{,}85 + B1*0{,}7)/\ DP$$

$$II = (A1 + A2*0{,}85 + B1*0{,}7 + B2*0{,}5 + B3*0{,}2 + B4*0{,}1 + B5*0{,}05)/\ DP$$

Nas expressões acima A_i, B_i denotam o número de publicações em periódicos qualificados como A_i e B_i respectivamente, e DP denota o número de docentes permanentes dos programas.

Realizamos a análise de dois cenários alternativos. No primeiro, levamos em consideração apenas os docentes permanentes internos do PPGCD. No segundo cenário, levamos em consideração o grupo todo, inclusive o docente permanente externo.

Conforme a Figura 2, observando apenas os dados relativos ao índice restrito, se o PPGCD fosse criado hoje, estaríamos em melhor posição que 34 programas de nível 3, 18 programas de nível 4, e 1 programa de nível 5, se consideramos o segundo cenário. Já no primeiro cenário (i.e., sem o docente externo), estaríamos em melhor posição do que 29 programas de nível 3, e que 5 programas de nível 4. Os resultados da análise considerando os dados relativos ao índice irrestrito são similares: perdemos apenas uma posição para UFCG no caso geral e duas posições para UFTPR e USP, se desconsiderarmos o docente externo.

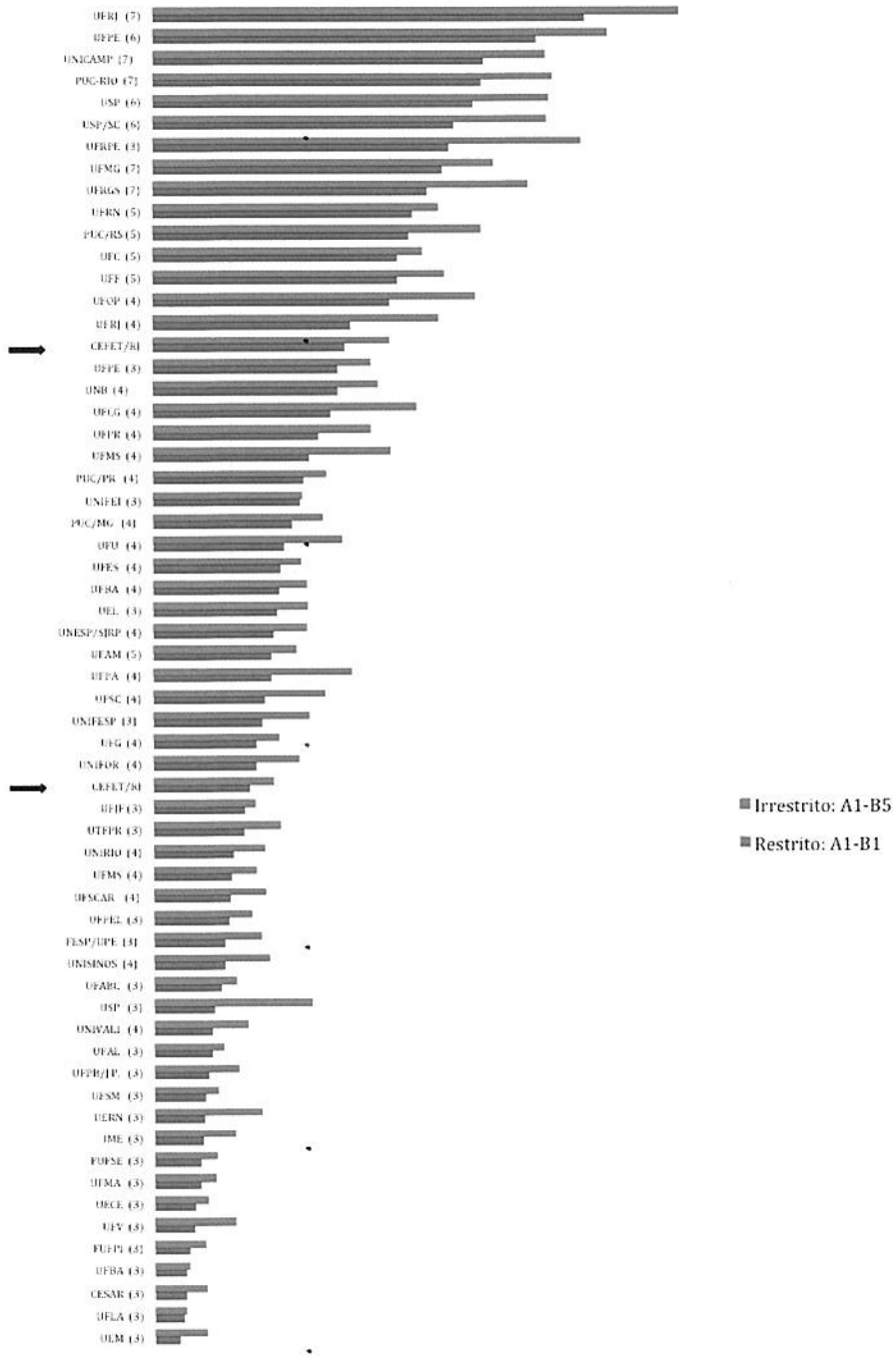

Figura 2 - Produção das IFES na Área de Computação

Referências

Bell, G., Hey, T., Szalay, A., (2009), "Beyond the Data Deluge", *Science*, v. 323, n. 5919 (Jun.), p. 1297–1298.

Berman, F., (2008), "Got Data?: A Guide to Data Preservation in the Information Age", *Commun. ACM*, v. 51, n. 12 (Dec.), p. 50–56.

Berriman, G. B., Deelman, E., Good, J., Jacob, J. C., Katz, D. S., Laity, A. C., Prince, T. A., Singh, G., Su, M.-H., (2007), "Generating Complex Astronomy Workflows", In: Taylor, I. J., Deelman, E., Gannon, D. B., Shields, M. [eds.] (eds), *Workflows for e-Science*, Springer London, p. 19–38.

CAPES, (2013), *Documento de Área da Ciência da Computação*, http://www.capes.gov.br/component/content/article?id=4656:ciencia-da-computacao.

Davenport, T. H., Patil, D. J., (2012), "Data scientist: the sexiest job of the 21st century", *Harvard Business Review*, v. 90, n. 10 (Oct.), p. 70–76, 128.

Deelman, E., Gannon, D., Shields, M., Taylor, I., (2009), "Workflows and e-Science: An overview of workflow system features and capabilities", *Future Generation Computer Systems*, v. 25, n. 5 (May.), p. 528–540.

Dhar, V., (2013), "Data Science and Prediction", *Commun. ACM*, v. 56, n. 12 (Dec.), p. 64–73.

DSC, (2014), *Data Science Central*, http://www.datasciencecentral.com/.

Han, J., Kamber, M., (2011), *Data Mining: Concepts and Techniques, Third Edition*. 3 edition ed. Burlington, MA, Morgan Kaufmann.

Jacobs, A., (2009), "The Pathologies of Big Data", *Commun. ACM*, v. 52, n. 8 (Aug.), p. 36–44.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., Shahabi, C., (2014), "Big data and its technical challenges", *Communications of the ACM*, v. 57, n. 7 (Jul.), p. 86–94.

Lazer, D., Kennedy, R., King, G., Vespignani, A., (2014), "Big data. The parable of Google Flu: traps in big data analysis", *Science (New York, N.Y.)*, v. 343, n. 6176 (Mar.), p. 1203–1205.

Liao, S.-H., Chu, P.-H., Hsiao, P.-Y., (2012), "Data mining techniques and applications – A decade review from 2000 to 2011", *Expert Systems with Applications*, v. 39, n. 12 (Sep.), p. 11303–11311.

Mattoso, M., Ocaña, K., Horta, F., Dias, J., Ogasawara, E., Silva, V., de Oliveira, D., Costa, F., Araújo, I., (2013), "User-steering of HPC Workflows: State-of-the-art and Future Directions". In: *Proceedings of the 2Nd ACM SIGMOD Workshop on Scalable Workflow Execution Engines and Technologies*, p. 4:1–4:6, New York, NY, USA.

NYU, (2014), *Data Science at New York University*, http://datascience.nyu.edu.

Ogasawara, E., Dias, J., Silva, V., Chirigati, F., de Oliveira, D., Porto, F., Valduriez, P., Mattoso, M., (2013), "Chiron: a parallel engine for algebraic scientific workflows", *Concurrency and Computation: Practice and Experience*, v. 25, n. 16 (Nov.), p. 2327–2341.

De Oliveira, D., Ogasawara, E., Baião, F., Mattoso, M., (2010), "SciCumulus: A Lightweight Cloud Middleware to Explore Many Task Computing Paradigm in Scientific Workflows". In: *2010 IEEE 3rd International Conference on Cloud Computing (CLOUD)*, p. 378–385

Stevens, R., Zhao, J., Goble, C., (2007), "Using provenance to manage knowledge of in silico experiments", *Briefings in Bioinformatics*, v. 8, n. 3 (May.), p. 183–194.

Wright, A., (2014), "Big Data Meets Big Science", *Commun. ACM*, v. 57, n. 7 (Jul.), p. 13–15.