Anexo I: Temas
Propostos 2023-2024

Este quadro indica algumas alternativas de professores(as) orientadores(as) e respectivos temas de interesse, porém não esgota as possibilidades. Candidatos(as) interessados(as) no PIBIC podem procurar diretamente os(as) professores(as) que desenvolvam pesquisas em temas de seu interesse.

FGV/EMAp 2023-2024

Aberto Paccanaro

Recomendando funções moleculares para proteínas

Os Recommender Systems são uma aplicação importante para aprendizado de máquina e inteligência artificial. Grandes empresas de tecnologia como Netflix, Google e Amazon dependem da recomendação de filmes, resultados de pesquisa ou produtos para seus usuários. Uma maneira muito eficaz de construir um recommender system é através da decomposição matricial, que aprende a recomendar itens aos usuários, reconstruindo uma matriz onde as linhas são os usuários, as colunas filmes ou produtos, e os valores representam interações entre usuários e itens.

Este projeto pretende explorar o uso de sistemas recomendados na biologia, para recomendar funções moleculares às proteínas. Para construir tal sistema, uma abordagem de decomposição matricial poderia ser usada empregando uma matriz onde as proteínas estão nas linhas, e as funções moleculares estão nas colunas (ou seja, aqui as proteínas assumirão o papel dos usuários, e as funções o papel dos filmes).

Estamos procurando um aluno apaixonado por inteligência artificial aplicada. O aluno implementará e avaliará o desempenho de um recommender system para previsão da função proteica. Eles devem ser capazes de ler e escrever código Python e também terão que aprender a executar o código de outras pessoas a partir do terminal.

O candidato selecionado atuará no PaccanaroLab (sala 534) ao lado de pesquisadores que trabalham na aplicação da inteligência artificial a múltiplas aplicações biológicas e médicas que incluem design de medicamentos, caracterização de doenças e sustentabilidade alimentar.

Comparando as características de representações de proteínas: modelo de linguagem versus Graph Neural Network.

Redes Neurais são muito boas em representar objetos no mundo real como vetores em espaço de alta dimensão. Essas representações podem então ser usadas em muitas aplicações diferentes, como reconhecimento de fala, tradução automática e visão

computacional. Aplicações em biologia e medicina envolvem a representação de proteínas como vetores.

Proteínas podem ser pensadas como longas sequências de letras representando aminoácidos. Isso naturalmente nos permite pegar modelos de deep learning que foram originalmente projetados para texto, e usá-los para representação proteica – este tipo de modelo é conhecido como um Modelo de Linguagem (LM). Por outro lado, existem grandes conjuntos de dados com informação sobre interações físicas proteína-proteína, que podem ser expressos como uma rede (chamada interatoma), onde os nós representam proteínas e os links representam uma possível interação entre eles. O interatoma pode ser usado para aprender representações proteicas por meio de técnicas como uma Graph Neural Network (GNN).

Acredita-se que o poder preditivo desses métodos venha de sua capacidade de codificar características de alto nível, tais como propriedades estruturais das proteínas. Isso faz sentido para os LMs porque a sequência em si contém informações sobre estrutura. Também faz sentido para os GNNs porque as proteínas interagem entre si por causa de sua estrutura 3D.

Estamos procurando um aluno interessado em trabalhar com métodos de deep learning. O aluno implementará e comparará espaços latentes gerados com LMs e GNNs, e determinará se está aprendendo características semelhantes. O aluno deve ser capaz de ler e escrever código Python, e terá que aprender a usar o terminal para executar o código de outras pessoas. O candidato atuará no PaccanaroLab (sala 534) ao lado de pesquisadores que trabalham na aplicação da inteligência artificial a múltiplas aplicações biológicas e médicas que incluem design de medicamentos, caracterização de doenças e sustentabilidade alimentar.

Avaliando a transferência das redes biológicas

Os dados gerados por experimentos biológicos de larga escala muitas vezes têm uma representação natural como redes – por exemplo redes de interação proteína-proteína, redes de interação genética, redes de co-expressão. Os experimentos necessários para gerar esses gráfos são muito caros e complexos, e várias abordagens de aprendizado de máquina foram desenvolvidas para prevê-los a partir de dados genômicos.

Em geral, essas abordagens baseiam-se na ideia de transferir partes de redes de organismos bem estudados (organismos modelos) onde a rede foi encontrada experimentalmente, para organismos menos estudados onde esses experimentos ainda não foram realizados.

É evidente que a qualidade das redes transferidas para um determinado organismo depende de quantos organismos evolutivamente próximos ele tem, e de quão bem conhecidas são as redes para esses organismos. No entanto, não temos como prever a qualidade da rede gerada por nossas abordagens de machine learning.

Este projeto pretende desenvolver uma abordagem de aprendizado de máquina para prever, para um determinado organismo, a qualidade de suas redes inferidas dependendo de suas relações evolutivas.

O aluno selecionado estará trabalhando no PaccanaroLab (sala 534) e terá a oportunidade de interagir com os demais membros do laboratório. O aluno poderá aproveitar uma série de pipelines já desenvolvidos no laboratório para prever redes.

Trata-se de um projeto de pesquisa e nosso objetivo é publicar seus resultados. Ele tem um forte componente de implementação e, portanto, um bom conhecimento de trabalho do Python é necessário e o aluno também terá que aprender a usar o terminal para executar o código de outras pessoas.

Métodos de integração de dados e de aprendizagem multi-perspectiva usando Fatoração de Matriz não Negativa.

Sistemas do mundo real são complexos e usualmente são estudados observando-se diferentes partes do sistema ou diferentes perspectivas do mesmo objeto. Por exemplo, muitos sistemas biológicos são estudados a partir de diferentes tipos de dados e experimentos (dados de sequenciamento genético, dados de expressão genética, etc), que descrevem níveis de informação diversos.

Em farmacologia, drogas podem ser representadas tanto usando características químicas, quanto seus efeitos colaterais ou as enfermidades que trata. As técnicas de Fatoração de Matriz não Negativa, inicialmente propostas para representar objetos num espaço de menor dimensão, tem demonstrado um bom desempenho para problemas de aprendizagem de máquina envolvendo a integração de diferentes tipos dados e múltiplas perspectivas de um mesmo objeto. A vantagem de se fazer essa integração, é que as diferentes perspectivas podem se complementar e, assim, é necessária analisá- las em conjunto para se ter uma visão sobre o todo. Apesar de haver abordagens bem sucedidas para esse problema, ainda não é claro qual é a melhor forma de se fazer essa integração para diferentes problemas de aprendizagem de máquina.

O objetivo deste trabalho é fazer uma comparação entre abordagens de Fatoração de Matriz não Negativa para integrar dados em diferentes aplicações. O processo inclui: uma revisão da literatura para conhecer os métodos de integração existentes para sistemas de recomendação e clustering, determinar os conjuntos de dados a ser usados para os experimentos (podendo ser artificiais), implementar os métodos para diferentes tarefas de aprendizagem de máquina, comparar os métodos existentes e, possivelmente, propor melhorias.

O aluno deve ser capaz de ler e escrever código Python e possivelmente Matlab. O candidato atuará no PaccanaroLab (sala 534) ao lado de pesquisadores que trabalham na

aplicação da inteligência artificial a múltiplas aplicações biológicas e médicas que incluem design de medicamentos, caracterização de doenças e sustentabilidade alimentar.

Alexandre Rademaker

Processamento de linguagem natural

Existem vários projetos interessantes para aplicações de técnicas de processamento de linguagem natural para problemas como:

1. O problema de detecção de implicação textual (text entailment, TE) é identificar quando duas sentenças (ou fragmentos de texto) estão relacionados de tal forma que a verdade de um fragmento de texto segue da verdade do outro.
2. Resposta automática à perguntas
3. Extração de Informações de textos

As 3 aplicações acima tem aplicações diretas na indústria e como tal, são de meu interesse imediato. Todas pode ser pensadas em contextos/domínios específicos ou gerais. Estas aplicações já foram parcialmente investigadas por alunos da EMAp orientados por mim e, desta forma, um continuidade natural de pesquisa pode ser conveniente.

Recursos léxicos para o Português

Relacionado ao tema anterior, quase todos os métodos de PLN demandam recursos lingúisticos, dados sobre a lingua natural a ser processada: informações sobre morfologia, gramática etc. Junto com colaboradores, mantenho alguns recursos para o Português de forma aberta.

A OpenWordnet-PT, ou simplesmente OWN-PT, é a WordNet de acesso aberto para o português. A OpenWN-PT está disponível em RDF/OWL e vem sendo expandida, melhorada e utilizada em projetos de processamento de linguagem nos últimos 10 anos. O dicionário morfológico para o Português chamado MorphoBr é outro importante recurso para processamento de textos.

Os corpora UD para o Português, em especial o GSD e o Bosque. A gramática do Português chamada PorGram.

Em comum, todos estes projetos demandam constante manutenção, vide issues abertos nos respectivos repositórios. Sejam bibliotecas de software, sejam interfaces de visualização ou consulta etc.

Exemplos de bibliotecas e ferramentas já desenvolvidas por alunos da EMAp e que ainda demandam expansão e revisão! Aos interessados em projetos aplicados e de programação:

- https://hackage.haskell.org/package/hs-conllu

- http://github.com/LR-POR/cl-conllu
- https://github.com/own-pt/wsi

Projetos de programação

Todas as propostas acima demandam algum interesse de programação. Dentre os paradigmas de programação mais influentes atualmetne, programação funcional detaca-se como uma forte tendência. Entre outros aspectos, a programação funcional tem um apelo especial para os alunos da EMAp, é o paradigma que talvez mais aproxime computação da matemática.

O desenvolvimento e/ou reimplementação de bibliotecas existentes em uma abordagem funcional é um excelente exercício para desenvolvimento de competência em programação funcional. Em todos os projetos acima citados, incentivo fortemente o uso de linguagens funcionais.

Além dos projetos já citados, algumas bibliotecas muito interessantes poderiam ser portadas para linguagens funcionais como Haskell ou Lean:

- https://pydelphin.readthedocs.io/
- http://wn.readthedocs.io

SUO-KIF translator to TPTP

Em projetos anteriores, começamos uma tradução do formato SUO-KIF para o formato TPTP. Esta tradução permite usarmos a ontologia SUMO em provadores automáticos de teoremas. Neste projeto, gostariamos de expandir esta transformação resolvendo bugs presentes na transformação atual e expandindo seu suporte para o formato TF0/TPTP. Como etapa seguinte, gostaríamos de usar SUMO em projetos de processamento de linguagem e contribuir com SUMO. A transformação inicial foi escrita em Lisp, mas esperasse a migração do código para Haskell. Outra direção possível de pesquisa é o reuso e possível reescrita do provador SNARK e seu suporte a procedural Attachments. Reimplementar um sistema como SNARK, embora pareça um projeto ambicioso, é uma excelenete oportunidade para desenvolver vários skills em programação.

Formalizing ALC SC/ND em Lean

Neste projeto, gostariamos de concluir a formalização dos sistemas dedutivos desenvolvidos para algumas lógicas de descrição em tese e provar as propriedades básicas destes sistemas. Código atual em https://github.com/arademaker/alc-lean.

Asla Medeiros e Sá

Visualização em bases de dados de biodiversidade

Coleções científicas de biodiversidade têm o compromisso de ser um registro permanente

da herança natural, constituídas de espécimes ou objetos relacionados ao seu domínio. Coleções digitais tipicamente contém uma versão digitalizada dos metadados correspondentes a cada item do inventário e podem, adicionalmente, conter arquivos multimídia tais como textos, registros fotográficos ou outros registros associados ao item, quando pertinente. Garantir a qualidade desses registros é uma tarefa complexa e de fundamental relevância. Fatores como o grande volume de dados e a interdependência entre múltiplas variáveis dificultam determinar até que ponto esses dados estão completos, corretos e se, de fato, fornecem uma boa cobertura geográfica, temporal e taxonômica das espécies correspondentes. O presente tema de trabalho propõe dar continuidade ao trabalho intitulado "Visualização de coleções científicas digitais de biodiversidade: um framework em Altair, Python."Defendido em março de 2021 no mestrado da EMAp (http://bibliotecadigital.fgv.br/dspace/handle/10438/30711)

Claudio Struchiner

COMORBUSS: simulador baseado em agente descrevendo a propagação de doenças de contato direto (ex. Covid-19) em uma população.

Principais características:

i) avaliação comparativa de quais serviços / locais mais contribuíram para a disseminação viral;
ii) quais estados sintomáticos (sintomático / assintomático / silencioso) se mostraram os vetores mais eficazes para a doença;
iii) rastreamento completo da rede de contatos;
iv) modelo de processo de diagnóstico e quarentena de pacientes com diagnóstico positivo ou sintomas graves;
v) modelos de distanciamento social;
vi) modelos de restrições parciais ou totais de atividades e serviços;
vii) liberdade na escolha de critérios para restringir e retomar as atividades de serviço, para simular e comparar diferentes medidas de contenção.

Informações adicionais sobre o programa podem ser obtidas em: https://comorbuss.org/Introduction

A seguinte nota técnica exemplifica o uso do programa para a simulação do impacto da reabertura das escolas na transmissão da Covid-19:

https://www.abrasco.org.br/site/wp-content/uploads/2021/05/Nota-Reabertura-Escolar-1.pdf

O candidato selecionado deverá mostrar proficiência em programação na linguagem

Python e terá a oportunidade de participar do desenvolvimento do simulador junto com a equipe de pesquisadores.

Dario Augusto Borges Oliveira

Aprendizado Profundo para Coleta de Evidências Locais de Mudança Climática e Clusterização de Eventos Extremos

A frequência, duração e intensidade de diferentes eventos climáticos extremos aumentaram à medida que o sistema climático se aquece. Por exemplo, a mudança climática leva a mais evaporação que pode agravar as secas e aumentar a frequência de fortes chuvas e nevascas. Esses eventos climáticos extremos geralmente resultam em condições ou impactos extremos, seja cruzando um limiar crítico em um sistema social, ecológico ou físico ou coocorrendo com outros eventos. Essas mudanças no clima são aparentes nos dados de EO (Earth Observation). Nesse contexto, espera-se uma demanda crescente por modelos resilientes que possam se adaptar aos efeitos do aquecimento global. Neste projeto, pretendemos explorar os efeitos do aquecimento global usando dados climáticos locais para a coleta de evidências de mudança no padrão de clima observado ao longo do tempo. Ademais, pretendemos realizar a clusterização das séries temporais, de forma a identificar padrões e tendências climáticas e, potencialmente, reconhecer eventos extremos.

An open set formulation for identifying new crop types from deep clustering using prototypes.

A aplicação de algoritmos de aprendizado profundo à observação da Terra EO nos últimos anos permitiu um progresso substancial em campos que dependem de dados de sensoriamento remoto. No entanto, dada a escala de dados em EO, criar grandes conjuntos de dados com anotações em nível de pixel por especialistas é caro, demorado e não raramente cria conjuntos de dados incompletos, ou com classes sub-representadas. Nesse contexto, priors são vistos como uma maneira atraente de aliviar o fardo da rotulagem manual ao treinar métodos de aprendizado profundo para EO, e permitem resolver problemas abertos de conjuntos, em algumas classes não são conhecidas. Este estudo propõe um método de agrupamento profundo on-line em que nem todas as classes são conhecidas, de modo a que se descubra novas classes através da análise dos agrupamentos. Como aplicação concreta, iremos validar a metodologia em dados de sensoriamento remoto de cultivo agrícola.

Uma Abordagem Multi-task para Identificação de Árvores Individuais em Florestas Tropicais usando Imagens LIDAR na Amazônia

Uma presença humana sustentável na Terra é uma das maiores preocupações da humanidade. Neste contexto o monitoramento eficiente de recursos naturais é imprescindível para que se consiga avaliar se as medidas de promoção de sustentabilidade são efetivas e suficientes. Neste projeto, planejamos monitorar a diversidade florestal usando dados de sensoriamento remoto e aprendizado profundo para

dar suporte ao planejamento de medidas de proteção e recuperação de áreas de floresta degradadas em larga escala;

O projeto irá utilizar dados de imagens LIDAR obtidas através de vôos não-tripulados em áreas da Amazonia que mapeiam com riqueza informações da superfície da floresta e permitem a delineação de copas de árvore e sua altura. O projeto irá então utilizar técnicas de aprendizado profundo, em especial de multi-task learning para identificar copas individuais de árvores, de modo a permitir a estimativa de população de árvores em uma determinada área de floresta. Em seguida o projeto ira analisar a população de arvores para extrapolar informações sobre biodiversidade na floresta.

Diego Parente Paiva Mesquita

How powerful are GFlownets?

Abstract:GFlownets foram propostas recentemente por Samy e Yoshua Bengio como uma alternativa à MCMC para espações discretos com estrutura aditiva (como grafos). Desde então, uma série de trabalhos se focou em utilizar GFlownets para descoberta causal, descoberta de novas drogas, otimização de caixa-preta, etc. No entanto, há pouca teoria sobre os limites de GFlownets e como diagnosticar situações patológicas no seu treinamento. Esse projeto tem como objetivo preencher esse buraco.

You must have a good grasp of: machine learning, graph theory
Skills you will develop: probabilistic machine learning.

Breaking down denoising diffusion probabilistic models
Abstract: Deep generative models leverage neural networks to approximate distributions over highly structured domains. Their major uses are i) estimating densities that can be used for anomaly detection; and ii) generating novel samples (e.g., see the first image here). Nonetheless, like most deep learning, little is regarding which components make these models thrive. For instance, an ICLR 2019 paper shows that variational autoencoders can be reduced to deterministic ones — with a few tricks on the side. In this project, we will disembowel denoising diffusion probabilistic models (DPPMs) to understand the core principles behind its success.

You must have a good grasp of: Machine learning principles, Programming
Skills you will develop: Deep learning, Probabilistic machine learning

GNNs sob medida

Abstract: Graph neural networks (GNNs) são o padrão de facto para diversas tarefas preditivas em grafos (e.g. prever interações em redes sociais ou propriedades moleculares). A direção comum na pesquisa de GNNs é tentar aumentar sua expressividade. Em outras palavras, há um foco em garantir que GNNs são capazes de

diferenciar grafos não-isomórfico.

Em contrapartida, o reflexo comum é um aumento no custo computacional de GNNs. Nesse projeto, focaremos na direção oposta. Mais especificamente, focaremos em desenvolver GNNs para classes restritas de grafos como os grafos bipartidos e as árvores. Nossa esperança é tomar inspiração em algoritmos especializados de isomorfismo de grafo para construir GNNs ótimas para essas classes e com custo computacional reduzido.

You must have a good grasp of: Machine Learning, Algorithms
Skills you will develop: Graph neural networks

Destilação e compressão de modelos Bayesianos

Abstract: Por vezes, é preciso distribuir modelos de machine learning em dispositivos de baixa capacidade computacional, como roteadores ou celulares. Nesse caso, é comum termos que comprimir grandes redes neurais devido a, e.g., limitações de memória. Uma maneira de fazer isso é a chamada destilação de modelo, na qual treinamos um modelo simples para imitar o comportamento de uma grande rede neural. Enquanto há uma vasta literatura sobre destilação de modelos estimados por métodos pontuais (e.g., máxima verossimilhança), há pouco sobre a destilação de modelos Bayesianos. Nesse caso, é razoável assumir que existe uma posteriori sobre grandes redes neurais, e gostaríamos de desenvolver um modelo mais simples (ou uma aproximação compacta da posteriori) que resulte nas mesmas distribuições preditivas. Nesse projeto, focaremos na destilação de redes neurais Bayesianas sem perder de foco seu uso final no apoio à tomada de decisão — i.e., que vamos usá-la para estimar funções de utilidade e/ou risco.

You must have a good grasp of: Machine Learning, Bayesian modeling

Skills you will develop: Probabilistic ML

Flávio Codeço Coelho

Projeto Infodengue:

O projeto Infodengue faz a modelagem de risco em tempo real da Dengue, Zika e Chikungunya. Este projeto se apoia em modelos matemáticos e estatísticos da transmissão destas arboviroses assim como em uma plataforma computacional desenvolvida para a agregação de dados epidemiológicos, Meteorológicos e de redes sociais, que alimentam a modelagem e oferecem a visualização interativa destes dados na web.

Atividades: Desenvolvimento Web usando a framework Django. Desenvolvimento e deploy de sistemas de captura distribuída de dados online.

Projeto Trajetórias:

Tem como objetivo produzir uma síntese da relação entre biodiversidade, serviços

ecossistêmicos e doenças transmitidas por vetores na Amazônia e desenvolver análises de cenários para áreas da região amazônica sob trajetórias concorrentes de uso da terra. Técnicas de modelagem matemática e estatística, formam a base da análise.

Atividades: Estruturação de um data lake baseado no software Apache Superset, para agregação das bases de dados utilizadas no projeto.
O estagiário terá a oportunidade de se aprofundar, em temas como linguagem SQL, uso avançado de Pandas, e Sistemas de informação geográfica(SIG) em combinação com a biblioteca geopandas.

Jorge Poco

Título: Urban Crime Analysis using Deep Learning

Métodos baseados em Deep Learning (DL) são o estado-da-arte em dezenas de aplicações, incluindo análise e percepção de crimes.

Este projeto visa investigar o uso de métodos baseado em DL para detecção e segmentação de objetos em cidades (e.g. árvores, estrutura em ruas, grafites) usando dados geoespaciais para soluções em planejamento urbano. Dois tipos de imagens têm sido coletadas, usando OpenStreetMap e dados geoespaciais. Imagens geoespaciais têm sido capturadas por satélites e nano-satélites, resultando em dados com resolução espacial e temporal para classificação de mapas de cidades a fim de se medir a probabilidade da incidência de crime, expansão urbana, e fluxo urbano. Neste projeto você será motivado a implementar modelos que usam módulos de atenção para comparar com métodos da literatura. Outros fatores econômico-sociais poderão ser considerados para percepção de criminalidade, assim dados desta natureza também estão sendo coletados.

Em um primeiro momento, daremos treinamento inicial sobre os métodos estado-da-arte em DL e GPUs para a/o candidata/candidato envolvendo o tópico central, depois compartilharemos códigos e material existente, bibliotecas usadas, e assim eles/elas devem ser capazes de codificar soluções aplicadas à dados geoespaciais adotando as bibliotecas mais reconhecidas do mercado e da academia. Dessa maneira, fluência em Python é esperado. A/O estudante pode ainda participar de publicações em curso, ou focar em desenvolver soluções para o seu portfólio profissional. Temos parcerias já estabelecidas com importantes players da indústria e do governo, exemplo, companhia privada de energia, e setores estratégicos em planejamento urbano. Tal candidata/o participará do grupo Visual Data Science - http://visualdslab.com, um espaço com extrema diversidade; com este projeto o candidato/a terá oportunidade de alavancar a sua carreira em Data Science, Visualização e machine Learning com soluções para aplicações reais.

Título: Explorando a mobilidade urbana

A mobilidade urbana, entendida como a movimentação de pessoas e bens dentro de áreas urbanas em diferentes modais, é um dos principais problemas das cidades modernas.

É comum o uso de tecnologias como monitoramento do tráfego por meio de câmeras e sensores, sistemas de controle de semáforos, sistemas de informação geográficos (GIS), aplicativos de mobilidade como serviço (Uber, Bike Rio, Moovit, Carpool), análise e coleta de dados em tempo real.

Além de ferramentas fornecidas pela administração pública, a participação da população é vital para enriquecer a quantidade e qualidade de informação a ser analisada. Aplicativos como Waze, permitem que os cidadãos contribuam de forma direta e indireta (crowdsourcing) com informação espaço-temporal sobre o estado do tráfego na cidade.
Este projeto visa fornecer insights que ajudem a otimizar a mobilidade urbana de grandes urbes, como Rio de Janeiro e São Paulo. Neste contexto serão criadas ferramentas interativas de visualização baseadas em algoritmos de aprendizado de máquina.

Realizaremos uma integração de dados heterogêneos provenientes de diversas fontes como repositórios públicos de dados abertos (Portal do Governo, Data.Rio) além dos dados informados por usuários de Waze e dados de sensores, semáforos, câmeras, obtidos por meio de parcerias. Para a gestão de dados de sensores, utilizaremos um conjunto de ferramentas e técnicas conhecidas como Internet das Coisas, que inclui protocolos de comunicação como MQTT e ferramentas de nuvem como Azure IoT.

Também extraímos padrões que nos ajudem a compreender o comportamento do trânsito nas suas diferentes modais, tais como algoritmos de agrupamento (clustering) para a identificar se um conjunto de alertas (trânsito lento, buracos na pista, alagamento) correspondem ao mesmo evento em uma mesma localidade ao longo do tempo.

Estamos à procura de um aluno que colabore nas tarefas de 1) integração de dados heterogêneos, 2) desenvolvimento de algoritmos de aprendizado de máquina para identificação de padrões (ex. eventos de trânsito a partir de alertas de Waze) e previsão de comportamentos. Serão utilizadas tecnologias de banco de dados (SQL, BigQuery) e tecnologias GIS e de visualização em ambiente Web (ex. Kepler.gl, D3.js)

O candidata/o participará do grupo Visual Data Science - http://visualdslab.com, um espaço com extrema diversidade; com este projeto o candidato/a terá oportunidade de alavancar a sua carreira em Data Science, Visualização e Machine Learning com soluções para aplicações reais.

Título: Estudiando la Dinámica de la Deforestación de la Amazonia.

La deforestación de la Amazonia y la degradación de la selva tropical son importantes

preocupaciones de los gobiernos brasileños, principalmente por su impacto en el medio ambiente y la agroindustria. Conocer estos fenómenos es esencial para establecer estrategias de control y promover técnicas alternativas de gestión forestal sostenible.

Este proyecto propone un enfoque basado en la ciencia de datos para abordar los problemas mencionados. El alumno participará en la construcción de herramientas analíticas capaces de revelar el comportamiento de los madereros, los lugares de tala ilegal y la dinámica de la deforestación y la degradación forestal a lo largo de los años.

También proponemos desarrollar métodos predictivos que ayuden a los organismos encargados de hacer cumplir la ley a identificar los lugares susceptibles de sufrir cambios en el uso del suelo, lo que permitiría anticiparse a la deforestación antes de que se produzca.

El trabajo se desarrollará con otros colaboradores de nuestro laboratorio y de la UFPE y USP. Además contaremos con la participación de ONG que proveerán la expertise del dominio.El candidato participará en el grupo de Visual Data Science - http://visualdslab.com, un espacio con extrema diversidad; con este proyecto el candidato tendrá la oportunidad de apalancar su carrera en Data Science, Visualização y Machine Learning con soluciones para aplicaciones reales.

Criação de visualizações em linguagem natural

Na análise exploratória de dados, baseia-se fortemente na representação visual dos dados para encontrar padrões, detectar outliers ou gerar hipóteses. Atualmente existem várias bibliotecas (por exemplo, matplotlib, D3, Altair, Plotly) que nos ajudam a criar visualizações com poucas linhas de código, porém, ainda é necessário saber programar para usá-las corretamente.

Por outro lado, temos os avanços do Processamento de Linguagem Natural, especialmente com técnicas de deep learning (por exemplo GPT-3, Transformers) que são capazes de realizar tarefas como criar poesia, música e até mesmo código para programar computadores. Embora os resultados ainda estejam na sua etapa inicial, existem vários aplicativos que podem ser criados para explorar seu potencial.

Neste projeto, queremos que o aluno investigue técnicas de PNL com aprendizado profundo para interpretar frases em linguagem natural para a criação de visualizações. Por exemplo, se o usuário digitar "mostre-me um gráfico de barras com precipitação mensal no eixo y ao longo do tempo", o sistema deve ser capaz de interpretar esse texto e convertê-lo em uma especificação de visualização que será posteriormente renderizada. (Veja o vídeo em https://bit.ly/3anxYtq).

Julio Cesar Chaves

Organização, curadoria e disposição de dados de pesquisas

Em termos práticos, cada aluno irá receber uma fonte de dados, um ambiente computacional de trabalho e a missão de organizar e documentar os dados, para que a consulta e o reuso sejam facilitados aos pesquisadores interessados.

Habilidades: banco de dados, modelagem informacional (sobretudo a parte da oficina).

Entregáveis: dados organizados, problemas de qualidade resolvidos e documentados, análises iniciais pertinentes ao contexto dos dados.

Redação de um artigo Data Descriptor, preferencialmente seguindo o modelo da revista Data-in-Brief (Data in Brief - Journal - Elsevier).

Projeto piloto de interface de navegação por voz sobre dossiês temáticos do acervo histórico do CPDOC através de uma skill da Alexa

A ideia é aproveitar os diversos dossiês temáticos existentes no portal do CPDOC para elaborar os primeiros roteiros dessa interação. Por exemplo, podemos armazenar na Alexa fatos sobre "A era Vargas", "Os anos JK", "A revolução de 1930", "O regime de 1964", O "AI-5", etc. Estes dossiês em geral comportam pequenos artigos e uma seleção de documentos textuais, audiovisuais, entrevistas e verbetes de dicionários histórico-biográficos. A ideia é abranger todos os temas, contudo sem entrar em detalhes. Caso o usuário deseje aprofundar, indicaremos a abertura do próprio portal.

Habilidades: lógica de programação, banco de dados e construção temática.

Entregáveis: a skill da Alexa (CPDOC FGV), e uma documentação sobre o trabalho desenvolvido.

Luiz Max

Tornando o DATASUS mais acessível

A crise da COVID-19 trouxe à tona a necessidade de dados públicos confiáveis e de fácil acesso para o auxílio à tomada de decisão.

Neste projeto o aluno vai utilizar a biblioteca PySUS para extrair dados das bases públicas do Ministério da Saúde e do IBGE para gerar tabelas e visualizações de interesse. Isso inclui dados de nascimentos, mortalidade, internações e atendimento ambulatorial. O objetivo final é disponibilizar uma série de bancos de dados curados e consolidados que

sejam de fácil consulta para profissionais de saúde e epidemiologistas.

Habilidades a serem desenvolvidas: Python, SQL (Big Query), visualização, Saúde Pública.

O espaço de árvores calibradas no tempo

Árvores filogenéticas ou filogenias são grafos planares utilizados para esboçar as relações evolutivas entre entidades biológicas. Uma das mais importantes aplicações das filogenias é no campo da filodinâmica, o estudo da dinâmica evolutiva de populações com evolução mensurável, como vírus e bactérias. Neste projeto o aluno vai construir rigorosamente o espaço de filogenias calibradas no tempo e então provar alguns resultados básicos sobre esse espaço, como: a cardinalidade do componente discreto, tamanho da vizinhança e propriedades de um passeio aleatório simples.

Habilidades a serem desenvolvidas: filogenética estatística, combinatória.

A conexão entre regras de pontuação próprias e o teorema da proporcionalidade de verossimilhanças

Regras de pontuação próprias (proper scoring rules, PSR) são dispositivos matemáticos para avaliar predições probabilísticas que incentivam a honestidade do analista ("forecaster"). São largamente utilizadas na avaliação de predições ("forecasts") de modelos em finanças, meteorologia e epidemiologia.

Gonçalves e Franklin (2019) recentemente provaram uma versão geral do chamado teorema da proporcionalidade de verossimilhanças (TPV), que diz que sob condições de regularidade, é sempre possível encontrar duas funções mensuráveis que representam a verossimilhança e diferem apenas por uma constante. Neste projeto, o aluno vai explorar a conexão entre as chamadas regras de pontuação próprias locais e o TPV, e investigar se é possível provar alguns resultados clássicos da teoria de PSR utilizando esse novo teorema.

Habilidades a serem desenvolvidas: estatística teórica.

Medidas de diagnóstico para MCMC com variáveis binárias

Métodos de cadeias de Markov Monte Carlo (MCMC) constituem uma classe extremamente útil de métodos numéricos, com aplicações em todas as áreas da estatística. A correta aplicação destes métodos depende, no entanto, depende de diagnosticar problemas de convergência e performance. Conquanto para variáveis contínuas existam várias medidas diagnósticas bem estabelecidas, para variáveis binárias ou categóricas o conjunto de ferramentas disponíveis é muito menor. Em trabalho recente, novas medidas foram propostas para variáveis binárias. Neste projeto o aluno irá estender alguns resultados já encontrados ao derivar as características assintóticas de algumas medidas bem como suas distribuições amostrais.

Habilidades a serem desenvolvidas: MCMC, R, métodos numéricos, cadeias de Markov de tempo discreto.

Processos gaussianos convexos e suas aplicações
Neste projeto estamos interessados em explorar certas propriedades das derivadas de processos gaussianos que permitem a modelagem de funções de interesse a partir das suas derivadas. Em particular, queremos estender resultados recentes que empregam emulação de funções para resolver os chamados problemas duplamente intratáveis. Uma vez desenvolvidos os métodos, vamos investigar uma aplicação nas chamadas "normalised power priors", que são muito utilizadas na análise de ensaios clínicos.

Habilidades a serem desenvolvidas: processos gaussianos; Stan/C++; estatística bayesiana.

Moacyr Alvim H. B. da Silva

Matemática no "Ranking" de Esportes.

Resumo: o objetivo é investigar, aplicar e comparar diferentes métodos de ranking e de previsão de esportes que fazem uso de ferramentas matemáticas variadas. Dentre as aplicações está prevista o ranking dos times do campeonato brasileiro e o ranking dos jogadores de tênis da década de 70 até hoje.

O que se espera dos alunos: um encontro semanal com o orientador é suficiente para acompanhamento das tarefas. Dentre as tarefas semanais espera-se que o aluno calcule o ranking dos times do campeonato brasileiro ao final de cada rodada para que os resultados sejam publicados no site da EMAp.

Rafael Pinho

Pet Analytics (Internet das Coisas + Machine Learning + Muitos Lambeijos) - Diagnóstico de Animais Domésticos, Ano 1.

Este é um projeto novo, que será conduzido em parceria com organizações de proteção animal como a Sociedade Zoófila Educativa (SOZED) e a ONG Os Indefesos.

Seu objetivo é o desenvolvimento de um dispositivo de Computação Vestível (Wearable Computing) - Peitoral H - para diagnóstico e monitoramento emocional de cães. É um projeto experimental, e buscaremos parcerias com cursos de veterinária do Rio de Janeiro para obter apoio especializado na definição e validação dos problemas que serão endereçados com Machine Learning e Visualização de Dados.

É obrigatório que o aluno se sinta confortável na presença de animais e tenha interesse de trabalhar diretamente com cães - nossos usuários adoráveis mas nem sempre colaborativos nos experimentos.

Health Analytics (Internet das Coisas + Machine Learning + Programação Web) - Plataforma de Monitoramento de Transporte de Cargas Pesadas, Ano 1.

Quando um motorista falha em cumprir as pausas planejadas em um transporte de longa distância, ou dirige em condições subótimas de atenção, os possíveis impactos para a sociedade não são aceitáveis. Alguns dos muitos desafios enfrentados por empresas de transporte de cargas pesadas são a garantia de que 1) o itinerário, que por exemplo inclui os descansos previstos em lei e a utilização de rotas seguras, é respeitado e 2) o motorista está em condições adequadas de direção.

Com base no protótipo da pesquisa "Health Analytics - Plataforma de Monitoramento Postural em Ambientes de Trabalho, Ano 1", estamos desenvolvendo o protótipo 1 de um assento veicular, para uso em veículos pesados no transporte de cargas entre países. O assento monitorará aspectos importantes das operações de logística para que sejam automatizadas ações de detecção e correção de inconformidades de itinerário e direção, em tempo real.

Para este projeto o aluno deve possuir conhecimentos de programação Neste projeto o aluno desenvolverá conhecimentos em Internet das Coisas, Programação Web, Análise de Dados e Visualização de Dados.

O aluno terá a oportunidade de desenvolver tecnologias e embarcá-las em caminhões de grande porte que percorrem todo o território nacional, interagir com o mercado e aprender sobre os desafios de dois setores importantes da economia - Logística e a Indústria Automotiva.

Health Analytics (Internet das Coisas + Machine Learning + Visualização de Dados) - Plataforma de Monitoramento Postural em Ambientes de Trabalho, Ano 2.

A pandemia do Coronavírus alterou drasticamente o ambiente de trabalho de um grande grupo de trabalhadores em todo o mundo, afetando as condições de ergonomia por um longo período de tempo.

O distanciamento social afastou a força de trabalho das mesas e cadeiras ergonomicamente planejadas dos nossos escritórios às camas improvisadas, sofás e mesas de jantar de nossas casas. Este período prolongado de más condições ergonômicas dificultou a capacidade de manter uma boa postura, agravando os desafios posturais do típico trabalhador de escritório que passa 15 horas sentado por dia e levando a um aumento da prevalência de dores lombares e no pescoço.

No primeiro ano de pesquisa utilizamos Internet das Coisas, Programação e Banco de Dados (e Corte e Costura!) para desenvolver uma plataforma de monitoramento de longo

prazo da postura de um usuário. Agora, no segundo ano da pesquisa, temos três objetivos:

1. Desenvolver painéis de dados e visualizações que serão utilizadas por profissionais de saúde - ortopedistas, fisioterapeutas, médicos do trabalho etc. - para apoiar o diagnóstico e o tratamento de doenças músculo-esqueléticas.

2. Refinar o protótipo inicial de Internet das Coisas, desenvolvendo uma segunda cadeira de escritório melhor preparada para a coleta e transmissão dos dados.

3. Utilizar Aprendizado de Máquina e desenvolver modelos que classificação de postura, qualidade de postura e reconhecimento de usuário.

Caso o aluno tenha interesse em trabalhar com a área de saúde mais diretamente, existe a proposta de migração da tecnologia desenvolvida para leitos hospitalares. Pacientes acamados enfrentam muitas dificuldades ergonômicas, algumas com sérias consequências para sua saúde após a internação. Representantes da UERJ e do Hospital Universitário Pedro Ernesto já manifestarem interesse em apoiar esta pesquisa.

Vincent Guigues

Inteligência artificial para os SAMUs

Resumo: O objetivo deste projeto e desenvolver um sistema de apoio as decisões de um SAMU. O projeto e dividido em 3 temas, cada um podendo servir de tópico de IC:

Algumas das ferramentas necessárias podem ser adquiridas no decorrer do projeto:

Alocação de ambulâncias as chamadas de emergência.

Escrevemos e implementamos em C++ um modelo de otimização para a alocação otimizada de ambulâncias a chamadas de emergência. O objetivo do projeto será implementar métodos para acelerar a resolução do problema e estudar e implementar estratégias de alocação alternativas. O módulo deverá comunicar com 2 outros módulos do projeto descritos a seguir.

Ferramentas: C++, bibliotecas de otimização, Python, Google API, banco de dados, Matlab, otimização, websocket.

Calibração de modelos para as chamadas de emergência.

O objetivo desta parte do projeto e propor, estudar e implementar modelos para as chamadas de emergência.

Ferramentas: R, C++, estatística.

Visualização das trajetórias das ambulâncias.

O objetivo e visualizar num Google Maps as trajetórias (discretizadas [tempo de discretização 5s]) das ambulâncias.

Ferramentas: banco de dados, C++, Google API, GraphQL, Hasura, Programação Web.

Yuri Saporito

Movimento Browniano e sua conexão com série de Fourier e Wavelets

Nesse projeto o aluno vai estudar a definição matemática rigorosa do movimento Browniano, o processo estocástico com maior impacto na matemática e ciência moderna. Para tanto, vamos ver, através da teoria de Fourier, a ideia de bases ortonormais e aproximação em espaços de funções. Uma vez que as ideias estejam sedimentadas, seguiremos para o mundo estocástico e estudaremos o movimento Browniano.

Teorema da Aproximação Universal

Redes neurais tem se mostrado uma ferramenta essencial nos dias de hoje. Nesse projeto, o aluno irá estudar as bases teóricas e históricas e provar algumas versões do Teorema da Aproximação Universal. Esse teorema mostra o poder das redes neurais no contexto de aproximação.

A Integral de Henstock–Kurzweil

A integral de Henstock–Kurzweil (HK) generaliza a integral de Riemann numa direção diferente da integral de Lebesgue. Nesse projeto, o aluno irá revisar as falhas da integração de Riemann (e sua história), entender a generalização de Lebesgue para então estudar a integral de HK.

Deep Galerkin Method (DGM) para solução de EDPs

Métodos de deep learning para solução de EDPs têm ganhado força na comunidade científica recentemente. Nesse projeto, o aluno irá implementar o DGM e outros métodos numéricos usuais para solução de EDPs para entender suas vantagens e desvantagens.