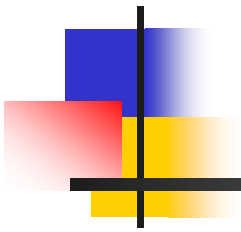


Data Science and Big Data



André C. P. L. F. de Carvalho



O mundo está sendo invadido



<https://proofofalien.com/will-aliens-attack-earth-did-nasa-confirm-the-alien-invasion/>

O mundo está sendo invadido



<https://www.em360tech.com/tech-news/tech-features/embrace-agile-environments-break-environment-management-bottleneck/>

© André de Carvalho - ICMC/USP



Apogeu da IA

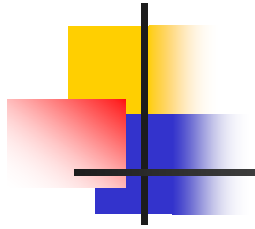
- Automação crescente
- Principal protagonista:
 - Aprendizado de Máquina (AM)
- Protagonismo de AM é influenciado por
 - Big Data
 - Ciência de Dados



Passado

Poucos dados





Presente

Inundação de dados



<https://inside.igneous.io/storing-files-beyond-big-data>



Inundação de dados

- Sem perceber, as pessoas geram dados a todo momento
 - Utilizam um cartão de fidelidade
 - Empresa aérea, supermercado, posto de combustíveis
 - Compram com cartão de débito ou crédito
 - Navegam na internet
 - Vão a um posto de saúde



Inundação de dados

- Dados gerados por empresas
 - De 2016 a 2020, é estimado que a geração de dados vai aumentar em 4.300%
 - Dados produzidos internamente
 - Atualmente, empresas usam apenas uma pequena parcela dos dados que produzem e armazenam
 - Empresas já sofrem para lidar com esses dados
- Além dos dados internos, haverá um grande aumento na quantidade de dados externos

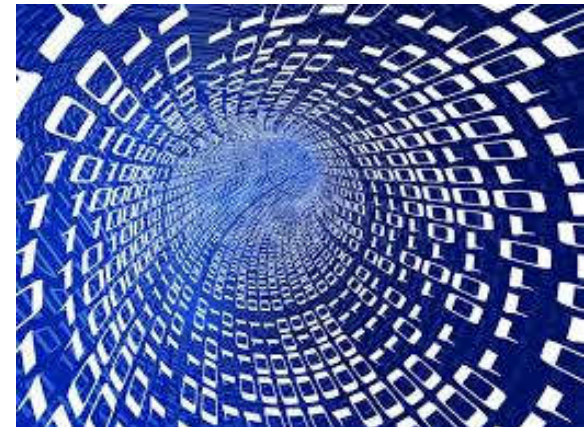


Dados externos

- São gerados por:
 - Outras empresas
 - Órgãos públicos
 - Organizações não-governamentais (ONGs)
 - Mídias
 - ...

Causas da inundação

- Avanços recentes nas tecnologias para
 - **Aquisição, armazenamento, transmissão e processamento** de dados
 - Maior **quantidade**, mais **rapidez** e menor **custo**

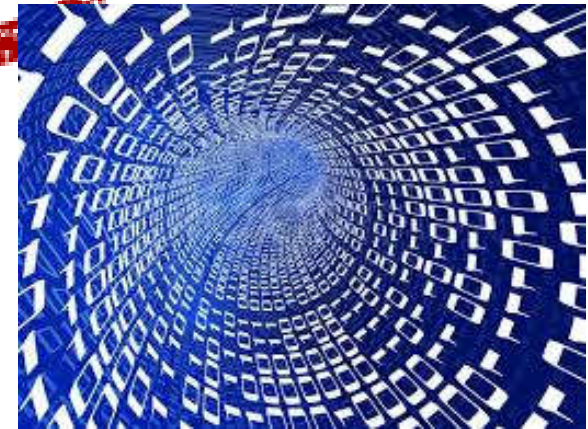




Causas da inundação

- Avanços recentes nas tecnologias para
 - **Aquisição, armazenamento, transmissão e processamento** de dados
 - Maior **quantidade**, mais **rapidez** e menor **custo**

Big Data





O que é Big Data?

**Incluído no Oxford English
em 2013**





O que é Big Data?

- Várias definições (40 líderes da área)
 - Dados que são grandes demais para sistemas tradicionais de processamento de dados



O que é Big Data?

- Várias definições (40 líderes da área)
 - Dados que são grandes demais para sistemas tradicionais de processamento de dados
 - Dados que precisam de novas técnicas para serem processados



O que é Big Data?

- Várias definições (40 líderes da área)
 - Dados que são grandes demais para sistemas tradicionais de processamento de dados
 - Dados que precisam de novas técnicas para serem processados
 - Dados que são muito complexos



O que é Big Data?

- Várias definições (40 líderes da área)
 - Dados que são grandes demais para sistemas tradicionais de processamento de dados
 - Dados que precisam de novas técnicas para serem processados
 - Dados que são muito complexos
 - Dados que são importantes



O que é Big Data?

- Várias definições (40 líderes da área)
 - Dados que são grandes demais para sistemas tradicionais de processamento de dados
 - Dados que precisam de novas técnicas para serem processados
 - Dados que são muito complexos
 - Dados que são importantes
 - Desafios e oportunidades decorrentes da disponibilidade de dados sobre tudo



Características de Big Data

- Grande **volume** de dados, gerados a uma grande **velocidade** e com uma grande **variedade** (3 Vs)
 - Volume: tanto de dados estruturados quanto de não estruturados
 - Variedade: vindos de fontes diferentes e que precisam ser integrados
 - Velocidade: gerados em fluxos cada vez mais rápidos

*Propostas por Doug Laney,
Consultoria Gartner, 2001*



Exploração de Big Data

- Dados gerados geralmente contêm informações relevantes
 - Uma vez analisados, podem trazer vários benefícios
 - Sociais, políticos e econômicos
 - Crescente interesse na análise de dados
 - Interesse em análise de dados não é uma atividade recente

O começo...

- Sumérios, Iraque, 4000 AC
 - Plantavam grãos em vale entre rios Tigre e Eufrates
 - Realizavam censos regulares para decidir quantidade de alimentos necessários
 - Registros do censo eram escritos em tábuas de argila
 - E depois analisados



O começo...

- Sumérios, Iraque, 4000 AC
 - Plantavam grãos em vale entre rios Tigre e Eufrates
 - Realizavam censos regulares para decidir quantidade de alimentos necessários
 - Registros do censo eram escritos em tábuas de argila
 - E depois analisados

Comer



Logo depois...

- No Egito, 2500 AC
 - Censos eram utilizados para decidir
 - Quantas pessoas seriam necessários para construir uma pirâmide
 - Como dividir a terra entre a população após as enchentes anuais do Nilo



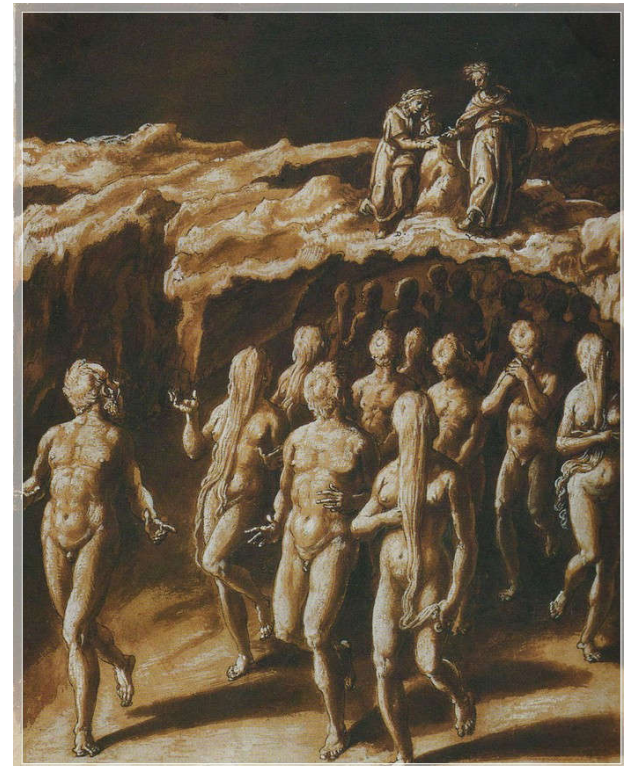


Século passado

- Grande expansão do uso de técnicas de análise de dados
- Como era:
 - Técnicas de análise de dados eram aplicadas a pequenos conjuntos de dados
 - Preparados pelos “donos dos dados”
 - Predições com baixa capacidade preditiva
 - “Modelos computacionais podem prever terremotos 10 anos antes (1971)”

Previsão

- Previsores humanos não tinham uma boa reputação
 - Divina Comédia de Dante
 - Poema (século 14)
 - Inferno
 - 8º Ciclo (Fraude)
 - Cartomantes
 - Adivinhadores
 - Astrólogos
 - Outros falso profetas





Pessoas podem errar previsões

- Baseadas na intuição
 - Os americanos precisam de telefones, nós não, pois temos muitos mensageiros (1878)
 - William Preece, Engenheiro chefe do serviço postal britânico
 - Nenhuma base de dados online substituirá meu jornal diário (1995)
 - Clifford Stoll, famoso astrônomo Americano
 - É impossível que o iPhone tenha mercado (2007)
 - Steve Ballmer, presidente da Microsoft























Neste século

- Mais conjuntos de dados
 - Maiores, mais complexos e dinâmicos
 - Big Data
- Novos desafios de análise de dados
- Necessidade de desenvolvimento de novas ferramentas
 - Motivaram avanços na análise de dados
 - Ciência de dados

~~Tempo~~ Dado é dinheiro

TOP 10 RANKING CHANGES SIGNIFICANTLY...

Only three brands that appeared in the BrandZ™ Global Top 10 in 2006—Google, Microsoft, and IBM—remain in the Top 10 in 2017.

	2006	Brand Value 2006 \$Mil.	2017	Brand Value 2017 \$Mil.
1	 Microsoft	62,039	 Google	245,581
2		55,834		234,671
3		41,406	 Microsoft	143,222
4	 中国移动 China Mobile	39,168	 amazon	139,286
5	 Marlboro	38,510	 facebook	129,800
6	 Walmart	37,567	 AT&T	115,112
7	 Google	37,445	 VISA	110,999
8	 IBM	36,084	 Tencent 腾讯	108,292
9	 citi	31,028	 IBM	102,088
10	 TOYOTA	30,201	 McDonald's	97,723

Source: Kantar Millward Brown / BrandZ™ (including data from Bloomberg)



~~Tempo~~ Dado é dinheiro

- Valor dos dados de 2 bilhões de perfis de usuários do facebook
 - Estimado em US\$ 32 bilhões em 2012 e US\$ 368 bilhões em 2016
 - Pelo menos 80 milhões dos perfis são falsos
- Valor global de vendas relacionadas a aplicações de Big Data
 - US\$ 7 bilhões em 2012 e US\$ 122 bilhões em 2015
 - Crescimento esperado para ~ US\$ 190 bilhões em 2019

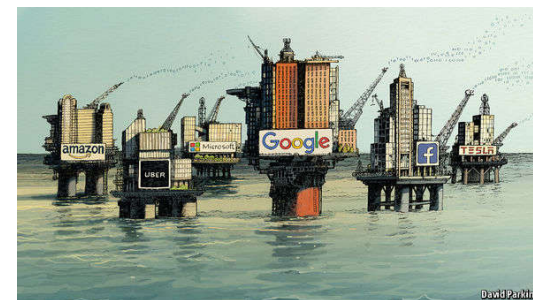


~~Tempo~~ Dado é dinheiro

- Em 2016, quase todo crescimento de arrecadação em propaganda nos EUA foi para Google e facebook
- Em 2017, 43% de todo o gasto em comércio eletrônico nos EUA foi para a Amazon
- Até março de 2017, Alphabet, Amazon, Apple, facebook e Microsoft lucraram mais de US\$ 25 bilhões

~~Tempo~~ Dado é dinheiro

- 1 tonelada de ouro vale US\$ 39,289,360.00
- Até hoje, 185.000 toneladas de ouro foram extraídos
 - Tudo somado: US\$ 7.4 trilhões (7400 000 000 000)
- Data G7 (Apple, Alibaba, Alphabet, Amazon, Facebook, Microsoft e Tencent) vale US\$ 3.7 trilhões (Julho de 2017)



Ciência de Dados

- Estuda princípios e técnicas para extrair conhecimento de um conjunto de dados
 - Novo, relevante e útil
- Pergunta chave da área:
 - Como extrair (de forma eficiente) conhecimentos em (grandes) conjuntos (fluxos) de dados





Big Data x Ciência de Dados

- Frequentemente usados como sinônimos
 - Big Data lida com tecnologias para coletar, gerenciar e processar (Big) dados
 - Ciência de Dados lida com criação de soluções para modelagem de dados
 - Capazes de extrair conhecimento de dados reais

Processar x Descobrir



Ciência de Dados

Tecnologias de
processamento
de dados
(ex. Big Data)



Ciência de dados

Tomada de decisão
orientada por dados



Ciência de Dados x Analítica

- Também usados como sinônimos
- Analítica (Analytics)
 - Processo científico de extrair conhecimento útil de dados para melhorar tomada de decisão
 - Tomada de decisão baseada em evidências
- Ciência de Dados engloba Analítica



O que é um Cientista de Dados

- Pessoa que é melhor em estatística que qualquer Engenheiro de Software e melhor Engenheiro de Software que qualquer Estatístico
 - Corolário de Priestley: Pessoa que explica melhor implicações gerenciais de resultados analíticos que qualquer Cientista e melhor em Ciência Analítica que qualquer MBA

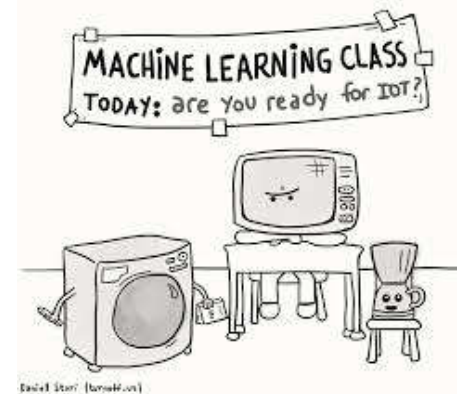


Ciência de Dados

- Técnicas tradicionais de análise de dados permitem apenas consultas simples
 - Quantos itens de um produto em particular foram vendidos em um dado dia?
 - Não conseguem responder consultas do tipo:
 - Que novo filme eu gostaria de assistir?
 - Dado o que estou sentindo, posso estar doente?
 - O que significa esse texto em chinês?
 - Por isso usamos Aprendizado de Máquina

Aprendizado de Máquina

- Computadores podem aprender a realizar uma tarefa, ao invés de serem explicitamente programados para isso
- Dezenas de milhares de algoritmos
 - Centenas de algoritmos novos a cada ano
 - Geralmente para superar algoritmos e
 - “Meu algoritmo é melhor que o seu”



Whose Button is Bigger?





AM *vs* computação tradicional

Escreve algoritmo
detalhando como
resolver um
problema

Computação
tradicional

Aprende a resolver
um problema
observando seus
dados

Aprendizado
de Máquina



AM *vs* computação tradicional

Pessoas programam

Computadores
programam

Computação
tradicional

Aprendizado
de Máquina



AM *vs* computação tradicional

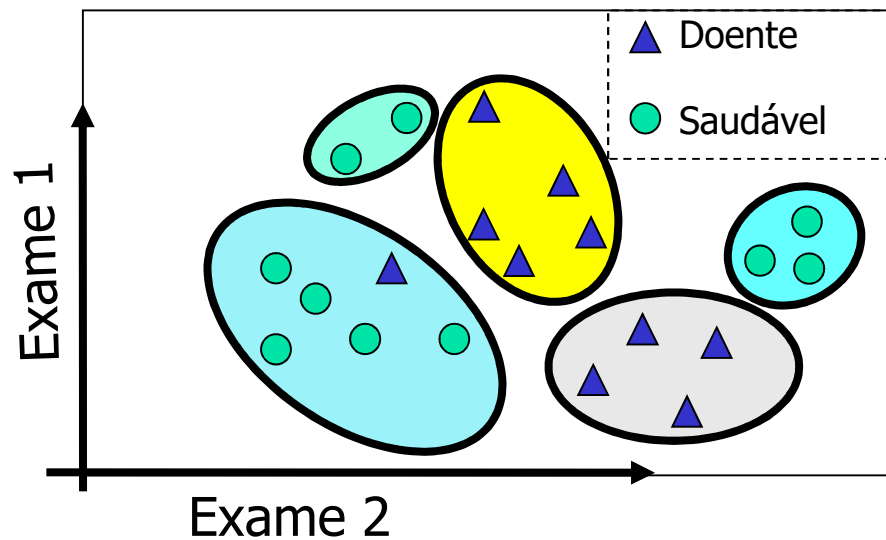
- Programação
 - Programa que funciona 90% das vezes é ruim
- Aprendizado de máquina
 - Modelo que acerta 90% das vezes pode ser o possível
 - E muito bom
 - Muitas vezes, é suficiente para ser útil



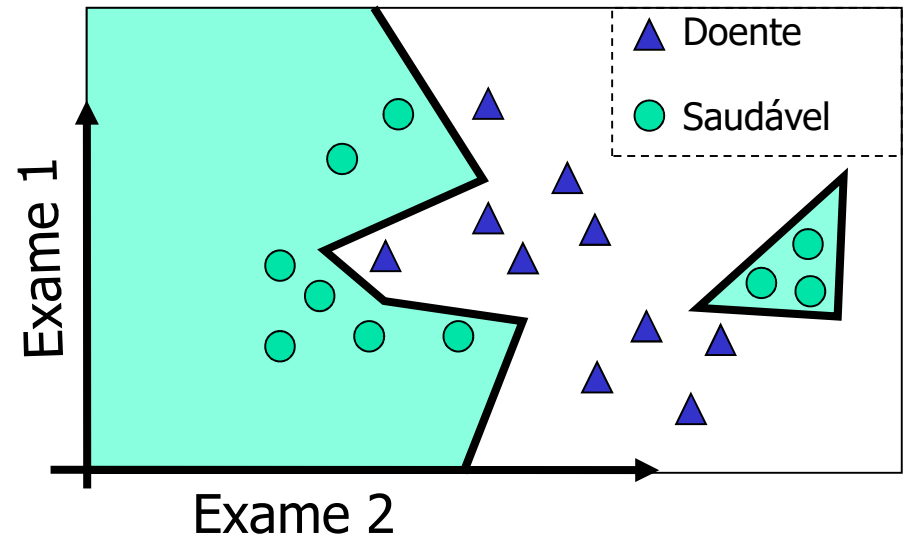
Aprendizado de Máquina

- Investiga técnicas capazes de aprender a resolver problemas
 - De forma automática, sem intervenção humana
- Bem sucedido em vários problemas reais de modelagem
 - Descritivos
 - Preditivos

Aprendizado de Máquina

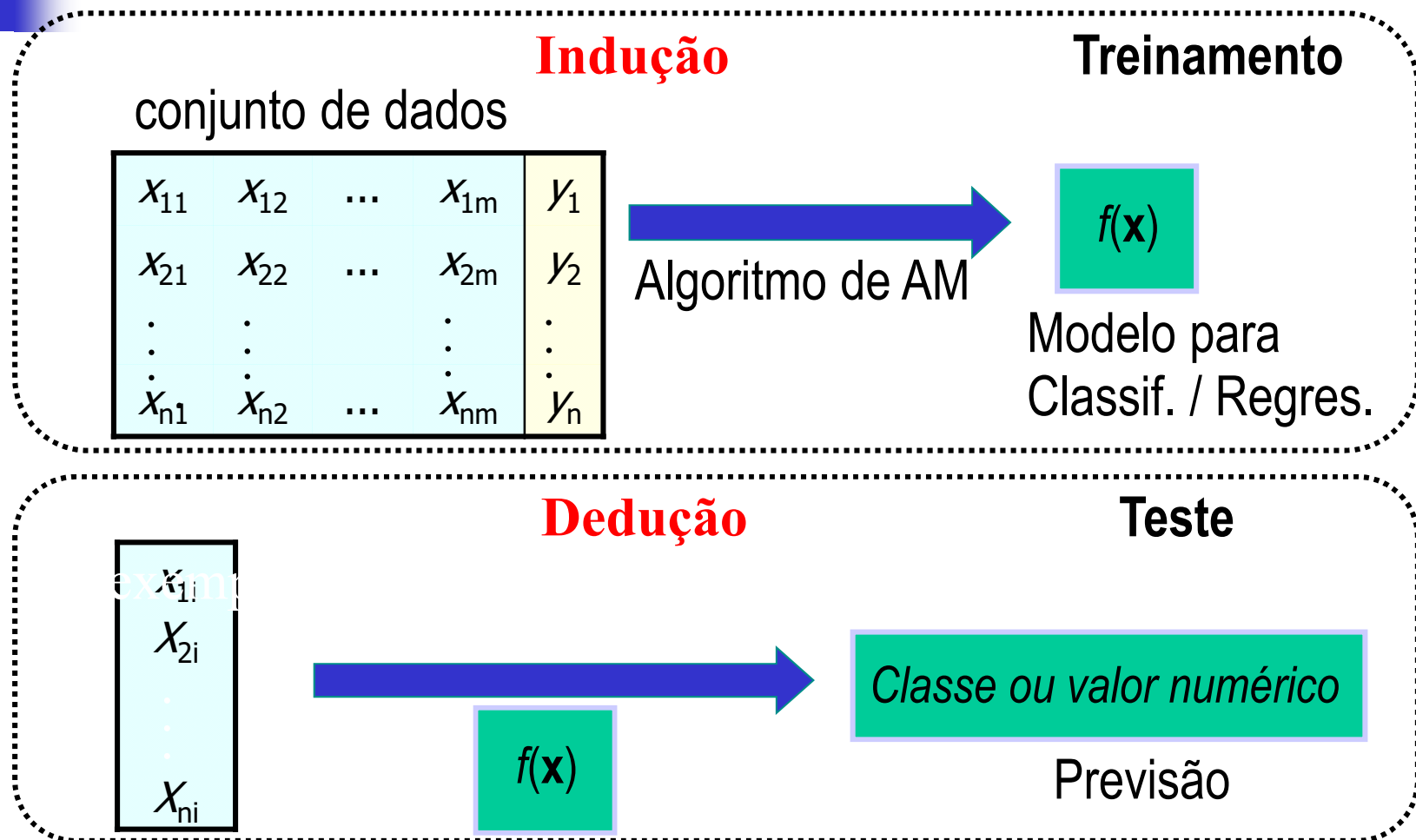


Descritiva
Agrupamento



Preditiva
Classificação

Indução e uso de modelos preditivos



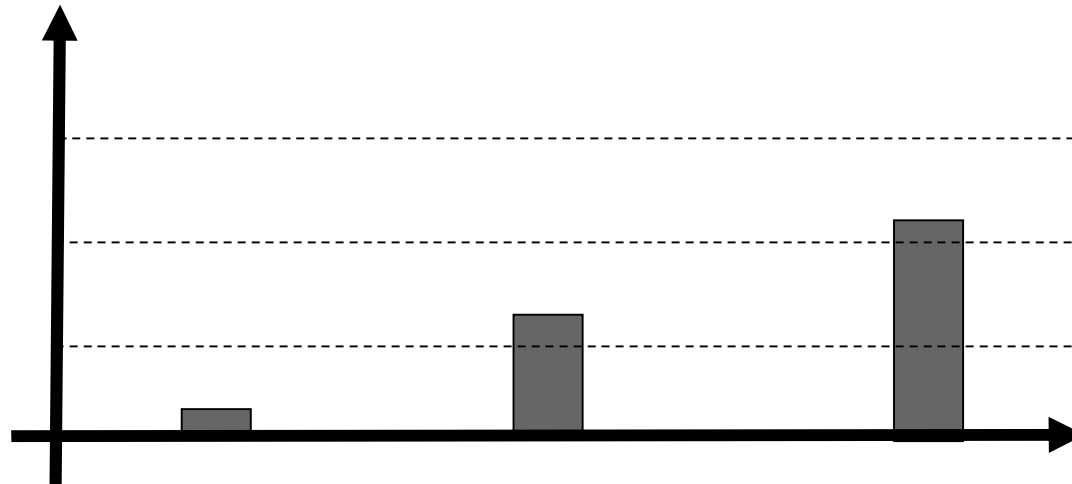


Reprodutibilidade

- Pesquisadores de IA buscam comparar novas soluções com soluções existentes
 - Com frequência, código de soluções existentes não estão disponíveis
 - Crise de reprodutibilidade
 - Semelhante a enfrentada por áreas como Medicina e Psicologia na década passada

Reprodutibilidade

- Pesquisadores de IA não compartilham seus códigos
 - Pesquisa com 400 algoritmos propostos nas 2 principais conferências de IA





Direito à explicação

- AM pode gerar modelos “caixa preta”
- Abril de 2016 parlamento europeu aprova a *General Data Protection Regulation* (GDPR)
 - Regulação geral de proteção de dados
 - Regula coleta, armazenamento e uso de informação pessoal
 - Artigo 22: *Automated individual decision-making, including profiling*
 - Tomada de decisão individual automatizada, incluindo criação de perfis



Direito à explicação

- Usuário pode pedir uma explicação para uma decisão sobre ele tomada por um modelo
- Oportunidades
 - Novos desafios para empresas
 - Novas oportunidades para pesquisadores
 - Novos algoritmos de AM e medidas de avaliação que
 - Possibilitem interpretação
 - Evitem discriminação



Proteção de privacidade

- *Fair Information Practices* (FIPs) for data
- Conjunto de 10 princípios para:
 - Coleta de dados
 - Acesso aos dados
 - Uso dos dados
 - Compartilhamento de dados



10 princípios das FIPs

1. Coleta limitada, legal e justa; com o consentimento ou conhecimento
2. Qualidade de dados, com dados relevantes, corretos e atuais
3. Especificação da finalidade no momento da coleta
4. Aviso de propósitos e direitos antes da coleta
5. Uso limitado (incluindo divulgação) aos fins especificados ou compatíveis



10 princípios das FIPs

6. Segurança por meio de barreiras efetivas
7. Clareza sobre como dados pessoais serão usados
8. Direito do indivíduo de acessar seus dados
9. Direito do indivíduo de modificar, completar e remover seus dados
10. Responsabilidade dos gerenciadores dos dados pela implementação



Ciência de Dados para o Bem

- Movimento sem fins lucrativos
 - Levar benefícios sociais para as pessoas e comunidades
 - Alguns programas são adotados por empresas
- Como ele ocorre?
 - Reuniões
 - Eventos
 - Estágios acadêmicos
 - Redes sociais





Ciência de Dados para o Bem

- Traz benefícios sociais para pessoas e comunidades
 - Bons serviços de saúde para todos
 - Desenvolvimento econômico de países pobres
 - Educação pública de qualidade
 - Energia limpa e barata
 - Melhor exercício da cidadania
 - Proteção ambiental
 - Meios de transportes mais seguros, rápidos e limpos



Decisões baseadas em evidências

- Movimento para que decisões sejam mais efetivas
 - Tragam mais benefícios
 - Reduzir influência de crenças e dogmas
 - Permitir políticas de estado
 - INGSA
 - Diplomacia científica

- International Network for Science Advice
 - *Science Government*
- Plataforma colaborativa para estimular a participação da ciência nas decisões públicas
 - Para isso, estimula organizações de assessoria científica e órgãos públicos a:
 - Usar evidências científicas
 - Discutir políticas públicas de estado
 - Capacitar interlocutores entre governo e academia
 - Desenvolver pesquisas no tema



- Disponibiliza fórum para formuladores de políticas, praticantes, academias e acadêmicos, permitindo:
 - Compartilhar experiências
 - Capacitar pessoas
 - Desenvolver abordagens teóricas e aplicadas
 - Para usar evidências científicas como política de informação em todos os níveis de governo



Estrutura do INGSA

- Coordenação
 - Prof. Sir Peter Gluckman, assessor de ciência do primeiro ministro da Nova Zelândia
- Capítulos
 - África
 - América Latina
 - Ásia
 - *Foreign Minister Science and Technology Advisor Network (FMSTAN)*
 - Rede de ministros de relações exteriores

Organizações parceiras



**INTERNATIONAL
COUNCIL
FOR SCIENCE**

Organização formada por 122 associações científica de, representando 140 países e por 31 uniões científicas internacionais

wellcometrust

Fundação de caridade cujo objetivo é melhorar a saúde



United Nations
Educational, Scientific and
Cultural Organization



International Development Research Centre
Centre de recherches pour le développement international

Canada



Centros de Ciência de Dados

- Columbia University, EUA
- Eindhoven University of Technology, Holanda
- Imperial College, Reino Unido
- Leiden University, Holanda
- New York University, EUA
- Tilburg University, Alemanha
- University of Edinburgh, Reino Unido
- University of Massachusetts at Amherst, EUA



Formação em Ciência de Dados

- Graduação, Mestrado e Doutorado
- Mais de 130 cursos de graduação
 - Eindhoven University of Technology, Holanda
 - Tilburg University, Alemanha
 - University of Nottingham, Inglaterra
 - University of Warwick, Inglaterra
 - University of Essex, Inglaterra
 - Yale University, EUA



Quem esta contratando CD

- Apple
- Booking.com
- Disney
- Google
- Greepeace
- Mercedes-Benz
- Red Bull F1



- Grande número de pesquisadores trabalhando em
 - Ciência de Dados
 - Aprendizado de Máquina
- Laboratórios
 - Analytics, LABIC, Estatística
- Formação de recursos humanos



NAP-AMDA

- Núcleo de Apoio a Pesquisa
Aprendizado de Máquina em Análise de
Dados
 - Centro de Pesquisa interdisciplinar
 - Mais de 60 pesquisadores
 - Universidade de São Paulo
 - Centros de Pesquisa e Universidades do Brasil e
do exterior

Colaborações nacionais

- CATI
- Embrapa
- IBM
- INPE
- Instituto Jardim Botânico
- PUC-Rio
- UFABC
- UFC
- UFF
- UFMG
- UFPE
- UFPR
- UFRJ
- UFRN
- UFSCar
- UFTPR
- UFU
- UNB
- UNESP
- UNIFESP
- UNICAMP



Colaborações internacionais

- Auckland University of Technology
- Arizona State University
- Central Queensland University
- East China Normal University
- Humboldt University
- Hong Kong Baptist University
- Kyushu Institute of Technology
- National Research Council of Canada
- Nanyang Technological University
- Norwegian University of Science and Technology (NTNU)
- The Ohio State University
- Technical University of Ostrava (VSB)
- Universidade do Porto
- Universidad de Salamanca
- Universidad Nacional de Rosario
- University of Alberta
- University of California, Riverside
- University of Kent
- University of Leipzig
- University of Lyon 2
- University of Pittsburgh
- University of Texas, Austin
- University of Regensburg
- University of Surrey
- Uppsala University





Áreas de interesse do NAP AMDA

- Aprendizado de máquina
- Ciência de dados
- Mineração de dados
- Pré-processamento de dados
- Agrupamento de dados
- Classificação de dados
- Metaheurísticas



Aplicados a

- Agricultura
- Bioinformática
- Ecologia
- Engenharia
- Finanças
- Medicina
- Petróleo e gás
- Redes sociais
- Sistemas de recomendação



CEPID CeMEAI

- Centro de Ciências Matemáticas Aplicadas à Indústria
 - ICMC-USP
- Início em 2013, duração de até 11 anos
 - Orçamento para primeiros 5 anos de 15 milhões de Reais da FAPESP
 - Mesma quantidade obtida de empresas
- Objetivo principal
 - Transferir conhecimento em computação, matemática e estatística para empresas



CEPID CeMEAI



Português

Search

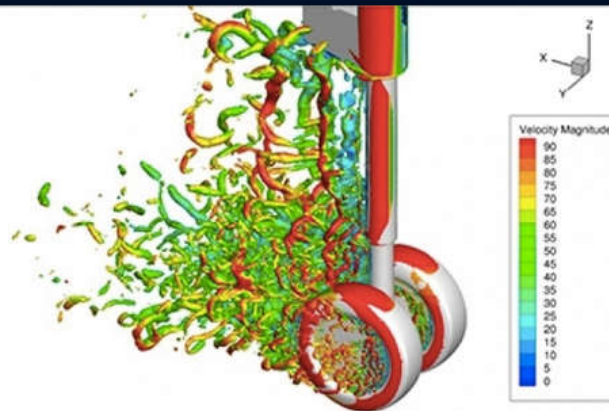
Center for Research in Mathematical Sciences Applied to Industry



About the Program

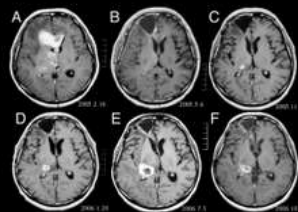
RIDCs

- Center for Metropolitan Studies
- Brazilian Research Institute for Neuroscience and Neurotechnology
- Center for Cell-Based Therapy
- Center for Computational Engineering and Sciences
- Center for Research and Development of Functional Materials
- Center for Research and Innovation in Biodiversity and Drug Discovery
- Center for Research in Mathematical Sciences Applied to Industry
- Center for Research on Inflammatory Diseases
- Center for Research on Redox Processes in Biomedicine
- Center for Research on Toxins, Immune-Response and Cell Signaling
- Center for Research, Teaching, and Innovation in Glass
- Center for the Study of Violence



Researchers analyze aircraft landing gear noise

A group at a Brazilian university, partnering with engineers from Boeing, are using computer simulations to identify sources of noise produced by aircraft undercarriages



CEPID 2013-2024

About the Center for Research in Mathematical Sciences Applied to Industry (CeMEAI)

WEBSITE: www.cemeai.icmc.usp.br

Principal Investigator: José Alberto Cuminato

Education and Knowledge Diffusion Coordinator: Lucio Tunes dos Santos

Technology Transfer Coordinator: Francisco Louzada Neto

HOST INSTITUTION

University of São Paulo (USP)
Av. do Trabalhador São-Carlense, 400
Caixa Postal 688
13586-590 - São Carlos, SP - Brasil
Tel.: (16) 3373-8159
contatocepid@icmc.usp.br
jacumina@icmc.usp.br

ASSOCIATED INSTITUTIONS

Institute of Aeronautics and Space (IAE)
Federal University of São Carlos (UFSCAR)
São Paulo State University (UNESP)
State University of Campinas (UNICAMP)

RESEARCHERS

Research projects supported in this Center at FAPESP's Virtual Library

Center leaflet (pdf):





Pesquisas no ICMC-USP

- **Bioinformática e Medicina**
 - Diagnóstico de câncer de câncer
 - Diagnóstico de demência
 - Prevenção de acidentes com idosos
 - Leiden University Medical Center)
- **Biometria e segurança**
 - Sequência de digitação e padrão de caminhada
 - Vigilância e detecção de falsificações
 - Polícia Federal e DARPA e UNICAMP



Pesquisas no ICMC-USP

- Ciência de Materiais
 - Projeto de novos materiais
 - UFSCar, Shell e Friedrich-Alexander-Universität Erlangen-Nürnberg (Alemanha)
- Finanças
 - Análise de crédito
 - Sistema automático de cobranças
 - Virgos IP Solutions
 - Detecção de fraudes
 - NEC (Alemanha)



Pesquisas no ICMC-USP

- **Processos industriais**
 - **Qualificação de madeiras**
 - Sguario Indústria de Madeira, Faber Castell e Universidade de Oulu (Finlândia)
 - **Otimização de alocação de peças cerâmicas em fornos industriais**
 - Optimized Decision Making - ODM
- **Meio ambiente**
 - **Modelagem de distribuição de espécies**
 - Instituto Florestal de São Paulo



Pesquisas no ICMC-USP

- Telecomunicações
 - AM para SAC
 - Grupo Algar
- Órgãos públicos
 - Recuperação de textos em livros antigos
 - Biblioteca Nacional da França Universidades de Bordeaux e Universidade de La Rochelle
 - Monitoramento de pacientes em UTIs
 - Hospital São João, Universidade do Porto



Conclusão

- Explosão de dados
- Big Data
- Ciência de Dados
- Crescimento da área
- Oportunidades na área
- Ciência de Dados para o bem



Conclusão

- Por que todo esse interesse?
 - Cada vez mais dados são gerados
 - Conhecimento precioso (vantagem competitiva) presente nesses dados
 - Avanços na capacidade de aquisição, armazenamento, processamento e transmissão de dados
 - Algoritmos cada vez mais eficientes para extrair conhecimento de dados



Conclusão

- Promessas exageradas
 - Cuidado com hypes
 - Muitas expectativas não vão se realizar
- Menos marketing, mais resultados
- Ainda é mais arte que ciência
- Utiliza conceitos de:
 - Computação
 - Aprendizado de Máquina
 - Estatística



Perguntas

