ACH5504

Mineração de Dados

Prof. Dr. Grzegorz Kowal

grzegorz.kowal@usp.br

https://sites.google.com/usp.br/ach5504

2° sem 2019 – segunda-feira, 14h00-18h00 – Edifício I1, sala 203

Introdução

- Quase tudo que vemos, lemos, ouvimos, escrevemos, medimos é coletado e disponibilizado em sistemas de informação computacionais.
- Vários fatores contribuem para o crescimento explosivo de dados. O problema tornou-se mais visível com a disseminação da Internet, em que indivíduos, empresas, governos, instituições não governamentais são produtores de conteúdo em potencial, transformando o mundo em uma enorme base de dados que é atualizada em tempo real, por milhares de pessoas, a cada segundo.

  Grandes Desafios da Pesquisa em Computação no Brasil, Sociedade Brasileira de Computação (SBC)

Motivação Econômica

- Muitos dados são coletados e armazenados
  - dados da web, comércio eletrônico
  - comprar em lojas e mercados
  - transações bancários e de cartão de crédito
- Computadores se tornaram mais baratos e mais poderosos
- Forte pressão competitiva entre empresas
- Fornecer serviços melhores e personalizados para diferenciar-se

Motivação Científica

- Mais dados coletados e armazenados
  - sensoriamento remoto de satélite
  - telescópios varrendo o espaço
  - dados de expressão de genes
  - simulações científicas
- Técnicas tradicionais não são apropriadas para analise de dados em tanta quantidade
- Boa parte dos dados não é analisada
- Analise de dados gera conhecimento

Mineração de Dados

- Como transformar **dados** em **conhecimento**?

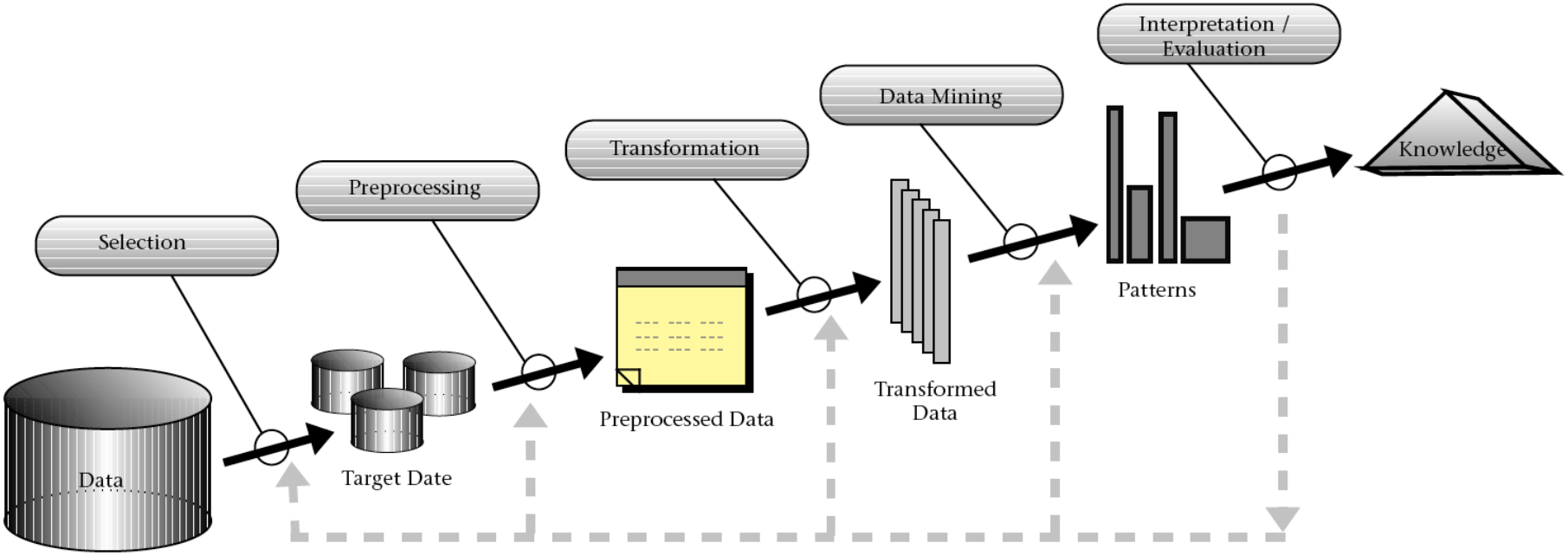
http://www.sliceofbi.com/2015/09/

Descoberta de Conhecimento em Base de Dados

- Knowledge Discovery in Database (KDD)
- “KDD is the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data“
- “Data mining [...] consists of applying data analysis and discovery algorithms that, under acceptable computational efficiency limitations, produce a particular enumeration of patterns (or models) over the data.”

  Data Mining to Knowledge Discovery in Databases, AI Magazine, Vol 17, No 3, 1996

Mineração de Dados

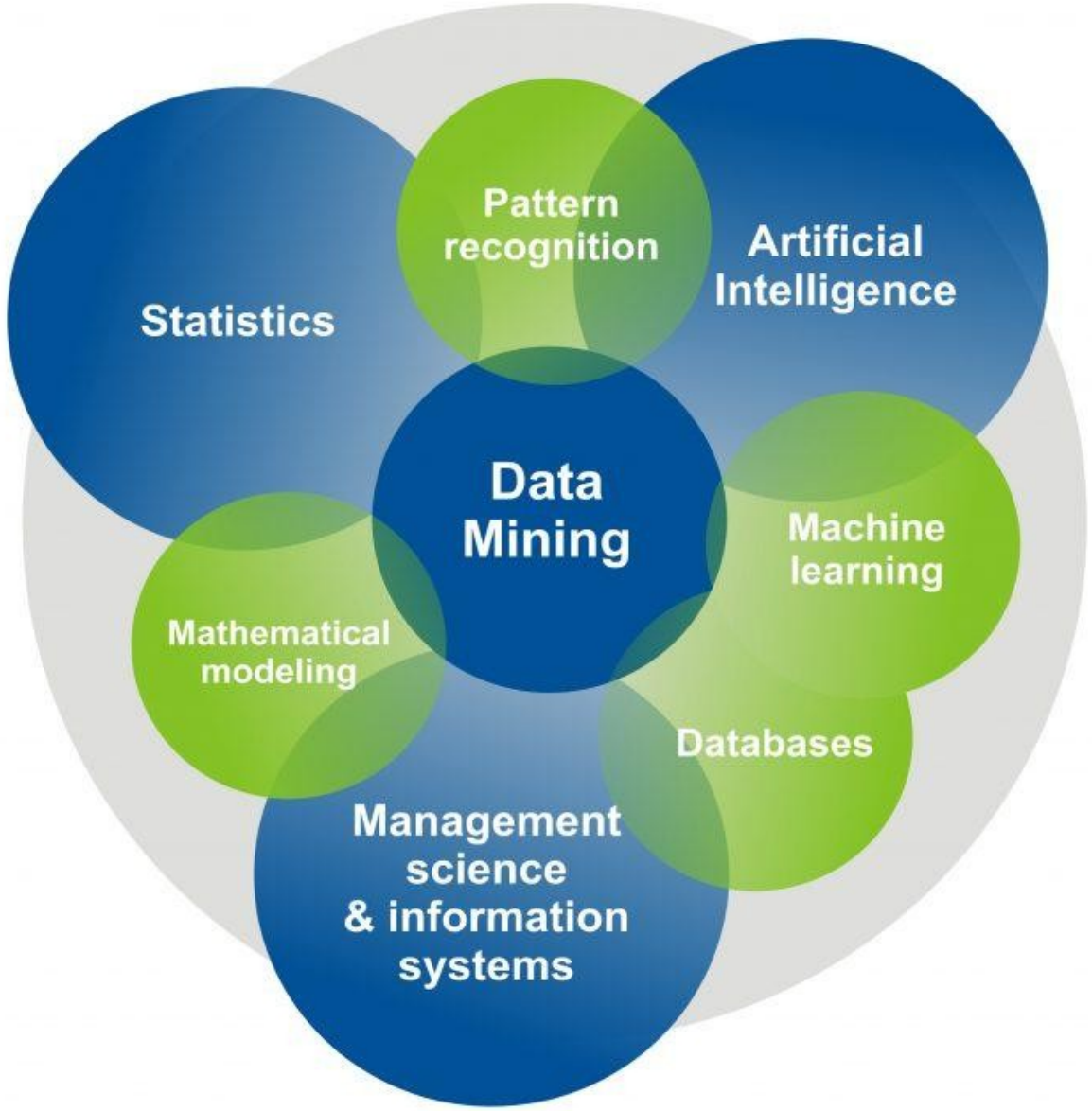


U. Fayyad, G. P.-Shapiro, and P. Smyth. From data mining to knowledge discovery in databases. AI Magazine, 17(3):37-54, Fall 1996

Mineração de Dados

- Interseção com várias áreas:
  - Inteligência Artificial
  - Aprendizado de Máquina (Machine Learning)
  - Reconhecimento de Padrões (Pattern Recognition)
  - Estatística
  - Banco de Dados (Database)
  - Modelagem Matemática

Mineração de Dados



Mineração de Dados

- Não é:
  - Fazer uma consulta a um banco de dados pelos dados de um cliente
  - Buscar documentos baseado na palavra “Brasil”
- Mas é:
  - Construir um modelo dos interesses do cliente para classificar itens como de seu interesse
  - Agrupar documentos com base em suas similaridades (praias do Brasil, futebol do Brasil,etc)

Mineração de Dados

- Tarefas Típicas:
  - **Preditivas**: usar algumas variáveis para predizer valores desconhecidos ou futuros de outras variáveis
  - **Descritivas**: encontrar padrões que descrevam os dados e sejam interpretáveis por humanos

Mineração de Dados

- Tarefas Típicas:
  - **Agrupamento**: ex. grupos de clientes com padrões de compra comum
  - **Classificação/Categorização**: ex. classificação biométrica de usuário
  - **Regressão**: ex. predição de demanda de um produto baseado em gasto com propaganda
  - **Regras de Associação/Dependência**: ex. quem compra A compra também B

Mineração de Dados

- **Dados** podem vir de várias fontes:
  - dados de satélite
  - compras na internet
  - localização pelo celular
  - postagens em redes sociais
  - amigos em rede sociais
  - vídeos na internet
  - fotos de viagens
  - notícias
  - e de onde mais?

Mineração de Dados

- Aplicações em Biotecnologia:

- **Genômica** - O Projeto Genoma, especialmente para o ser humano, levou um longo tempo de pesquisa e apoio mundial para identificar os mais de 20.000 genes e a sequência de todos os 3 bilhões de bases genômicas. Este projeto custa bilhões de dólares globalmente, mas as empresas de biotecnologia de hoje usam o banco de dados Big Data, que pode decodificar genomas inteiros por apenas milhares de dólares. O mercado de genômica ajuda diferentes empresas de dados que usam frameworks e ferramentas para realizar tarefas de computação enormes e complicadas para analisar dados genéticos, médicos e biológicos. Essas empresas geralmente trabalham com gigantes de hardware para melhorar o desempenho de seus aplicativos e seus resultados de análise de Big Data.
- **Agricultura** - Big data também pode ser uma aplicação importante no campo da agricultura. Os dados recolhidos da tecnologia GPS são armazenados no âmbito do Big Data e vários tratores com GPS podem ajudar os agricultores a lidar com as mudanças ambientais, implementando a agricultura com precisão. A análise de dados também está mudando o panorama da indústria de biotecnologia, com sua contribuição para a pesquisa genética na criação de organismos geneticamente modificados. Essas culturas modificadas podem ser modificadas com insumos da coleta de dados do Big Data para melhorar o rendimento da colheita, sobreviver à mudança de condição e obter plantas livres de doenças.

Fonte: https://explorebiotech.com/applications-of-big-data-in-biotechnology/

Mineração de Dados

- Aplicações em Biotecnologia:

- **Automação Farmacêutica** - De acordo com quase todas as empresas farmacêuticas, ela recebe milhões de compostos antes de se apropriar aos ensaios pré-clínicos. Para a jornada para a descoberta bem sucedida de medicamentos, consome uma enorme quantidade de tempo e dinheiro. Portanto, existem muitas ferramentas de software que ajudam na eficiência e menos tempo para a descoberta de medicamentos. A modelagem baseada em Big Data usa tamanho e armazenamento grandes, como terabytes de dados e informações de diferentes compostos e suas características. Por isso, atua como uma biblioteca virtual que possui informações de milhões de compostos para identificar os compostos que provavelmente terão sucesso. Esses programas de modelagem preditiva comparam os critérios do estudo e os resultados desejados com a doença alvo e as estruturas químicas. A automação farmacêutica reduz os riscos, economiza dinheiro e oferece ciclos mais rápidos de pesquisa no mercado.

- **Cuidados de saúde** - Tecnicamente, o setor de saúde da biotecnologia ficou para trás do que outros no uso do banco de dados de Big Data. As partes interessadas do setor de saúde agora têm acesso a novos e promissores segmentos de conhecimento. Essas informações na forma de Big Data fornecem complexidade, diversidade e cronogramas. A indústria farmacêutica exporta, paga e fornece análises de big data para obter insights. Com esses avanços tecnológicos na indústria de biotecnologia melhoraram. Sua capacidade de trabalhar com esses dados, embora os arquivos sejam enormes e tenham uma organização de banco de dados diferente, o que aumenta a condição e a taxa de desenvolvimento da assistência médica farmacêutica.

Fonte: https://explorebiotech.com/applications-of-big-data-in-biotechnology/

Mineração de Dados

- Aplicações em Biotecnologia:

- **Crowdsourcing** - De acordo com a Wikipedia, crowdsourcing é o modelo de sourcing em que indivíduos ou organizações obtêm bens e serviços. Esses serviços incluem idéias e finanças que formam um grupo grande, relativamente aberto e em rápida evolução de usuários da Internet. Ele divide o trabalho entre os participantes para obter um resultado cumulativo. Por isso, é comumente usado terceirização de projetos trabalhistas e empresariais. Algumas empresas farmacêuticas criaram plataformas de jogos on-line que envolvem perfis de doenças, desafios de pesquisa e soluções para desafios médicos. Com o crowdsourcing, diferentes pacientes conduziram pesquisas por meio de pesquisas on-line que capacitam o consumidor a conduzir seus próprios estudos e pesquisas, carregar seus próprios dados médicos e contribuir com conhecimento sobre sua condição e sintomas para beneficiar toda a comunidade médica.

- **Desenvolvimento de negócios** - Todos os dias, o corpo de informações sobre descobertas científicas e progresso farmacêutico vem de diferentes fontes, apresentando uma enorme quantidade de dados para as indústrias biofarmacêuticas a fim de encontrar potenciais oportunidades de licenciamento. Algumas grandes empresas de biotecnologia e biotecnologia recorreram a tecnologias de análise e mineração de dados para vasculhar diferentes fontes de Big Data e fornecer as informações exatas que buscam. Esse Big Data sempre cresce e se desenvolve junto com o negócio. Portanto, o Big Data aumenta a receita e o lucro total do negócio e, assim, desenvolve os negócios da biotecnologia.

Fonte: https://explorebiotech.com/applications-of-big-data-in-biotechnology/

Mineração de Dados

- Aplicações em Biotecnologia:

- **Análise de sentimentos** - Entre as ferramentas do Big Data, a análise de sentimentos é uma ferramenta que ajuda a analisar as postagens e os comentários das redes sociais. As organizações usam principalmente para pesquisa de marketing, publicidade e relações públicas. Por exemplo, muitas empresas o usam para encontrar a reação do consumidor e obter seu feedback. No entanto, as plataformas de mídia social contêm milhões de comentários relacionados à saúde porque os consumidores de serviços de saúde estão compartilhando informações pessoais e públicas sobre doenças e condições médicas. Algumas empresas estão criando um grupo online e uma comunidade para centralizar e descobrir novas descobertas e tecnologias. Quando usados em conjunto com o crowdsourcing, essas ferramentas fornecem fontes de trabalho livre e informações infinitas.
- **Prevenção da Fraude das Drogas** - Todos os dias, em países em desenvolvimento, drogas falsas matam muitas pessoas e afetam sua condição de saúde. Devido a esse triste cenário, pacientes e familiares perderam a esperança de diferentes empresas farmacêuticas e perderam vendas. A Organização Mundial da Saúde estima que 700.000 pacientes na África perecem como resultado de versões falsas de remédios contra a malária e a tuberculose, e o problema custa aos fabricantes de medicamentos US $75 bilhões por ano. O problema matador levou a startup Sproxil a trabalhar com a gigante de tecnologia IBM (IBM) para permitir que as empresas farmacêuticas analisassem fontes de Big Data para detectar padrões de atividade de medicamentos falsificados. O Sproxil pretende acumular grandes quantidades de dados transacionais com um sistema que permite que os pacientes enviem códigos de mensagem de texto de frascos de medicamentos para saber se os medicamentos são autênticos. Com a tecnologia de visualização da IBM e outras análises, os fabricantes de medicamentos podem acessar uma grande quantidade de dados sobre transações de medicamentos em tempo real, de acordo com a Big Blue. Presumivelmente, as fraudes com medicamentos podem ser vistas.

Fonte: https://explorebiotech.com/applications-of-big-data-in-biotechnology/

Objetivos

O curso tem como objetivo ensinar a explorar e analisar grandes quantidades de dados biológicos, por meios automáticos ou semiautomáticos que promovam a descoberta de padrões, a interpretação do conhecimento extraído e a visualização científica.

Conteúdo Resumido

- As origens da mineração de dados e seus desafios atuais.
- Tarefas de mineração de dados: classificação, agrupamento, associação, predição, regressão, etc.
- Tipos e qualidade de dados.
- Pré-processamento e técnicas de mineração de dados: medidas de similaridade e dissimilaridade
- Aprendizagem de máquina.
- Mineração de dados complexos, dados espaciais e outros tipos.
- Validação de resultados.
- Aplicações de mineração de dados e impacto social.
- Processo de Descoberta de Conhecimento em Bases de Dados (KDD).

Ferramenta

- Interpretador de Python (Online):
  - https://repl.it/
- Interpretador de Python (Offline):
  - Para Linux (Ubuntu)/Mac: python
  - Para Windows: https://www.python.org/downloads/windows/
- Bibliotecas:
    - Numpy (pacote para vetores e matrizes): https://www.numpy.org/
    - Matplotlib (gráficos) https://matplotlib.org/
    - BioPython (ferramentas para bioinformática): https://biopython.org/
  - Tutoriais, ex. https://www.springboard.com/blog/data-mining-python-tutorial/

Avaliação

- **Método**: Exposição teórica, seguida de exercícios, seminários e trabalhos práticos com o uso de computador.
- **Critério**: Serão atribuídas notas aos trabalhos práticos e seminários, e serão propostas provas em sala de aula. A nota final será calculada pela média ponderada dessas notas.
- **Norma de recuperação**: Uma prova escrita/prática dentro do prazo regimental e um trabalho. A nota da primeira avaliação será a média aritmética das provas e do trabalho. A nota da segunda avaliação será a média aritmética entre a nota da prova de recuperação e a nota final da primeira avaliação. O aluno será aprovado se obtiver nota na segunda avaliação igual ou superior a 5,0 (cinco).

Bibliografia

- **Alpaydin, E.**, *Introduction to Machine Learning*, MIT Press, 2004
- **Flach, P.**, *Machine Learning:The Art and Science of Algorithms that Make Sense of Data,* Cambridge University Press, 2012
- **Mitchell, T. M.**, *Machine Learning*, McGraw-Hill, 1997
- **Tan, P., Steinbach; M. E., Kumar, V.**, *Introduction to Data Mining*, Addison Wesley, 2006
- **Witten, I. H.; Frank, E.**, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edition. Morgan Kaufmann, 2005.

Cronograma

05/08 – semana de recepção de alunos
12/08 – apresentação da disciplina, introdução a mineração de dados
19/08 – origens e desafios da mineração de dados, tarefas de mineração de dados
26/08 – tipos de dados, qualidade de dados
02/09 – Semana da Pátria - não haverá aula
09/09 – medidas de similaridade e dissimilaridade, pré-processamento de dados
16/09 – classificação: preliminares, abordagem geral para a resolução de um problema de classificação
23/09 – regras de associação, geração de conjuntos de itens frequentes e regras
30/09 – **prova escrita**
07/10 – análise de agrupamentos: algoritmo k-médias, agrupamento hierárquico, avaliação de agrupamentos
14/10 – aprendizagem de maquina, redes neurais artificiais
21/10 – mineração de dados complexos, espaciais, texto, e outros tipos de dados
28/10 – Consagração ao Funcionário Público - não haverá aula
04/11 – aplicações de mineração de dados, tendências em mineração de dados
11/11 – Processo de Descoberta de Conhecimento em Bases de Dados (KDD): Limpeza dos dados; integração; transformação; redução de dimensionalidade.
18/11 – resolução de problemas relacionados a trabalhos
25/11 – entrega de trabalhos, revisão geral
02/12 – **prova substitutiva**