# Mineração de Dados em Biologia Molecular

## Ciência de Dados

Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

## Tópicos

- Introdução
- Análise de dados
- Ciência de dados
- Crescimento da área
- Ciência de dados para o bem



## Introdução

- Dados geralmente contém informações relevantes
  - Uma vez analisados, podem trazer vários benefícios
  - Cada vez mais analisados por entidades públicas e privadas
  - Análise de dados não é uma atividade recente



## O começo...

- 4000 A.C., Babilônia
  - Censos regulares eram realizados
    - Decidir quantidade de alimentos que seria necessário encontrar para alimentar toda a população
    - Registros do censo eram escritos em placas de barro



## O começo...

- 4000 A.C., Babilônia
  - Censos regulares eram realizados
    - Decidir quantidade de alimentos que seria necessário encontrar para alimentar toda a população
    - Registros do censo eram escritos em placas de barro




## O começo...

- 2500 A.C., Egito
  - Censos eram utilizados para decidir
    - Quantas pessoas seriam necessários para construir pirâmides
    - Como dividir a terra entre a sua população após as enchentes anuais do rio Nilo

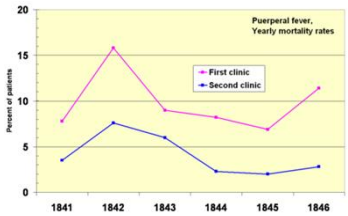
## Mais recentemente

- 1847, médico húngaro Ignaz Semmelweis
  - Começou a trabalhar no hospital geral de Viena em 1846
    - Supervisão de partos difíceis
    - Ensino de obstetrícia
    - Tomar conta dos registros médicos
      - Chamou sua atenção diferentes taxas de morte por febre puerperal em duas clínicas do hospital (1841-1846)



## Mais recentemente

- Taxas diferentes de morte por febre puerperal em 2 clínicas do hospital (1841-1846)






## Mais recentemente

- Febre puerperal
  - Comum em hospitais em meados do século XIX e frequentemente fatal
- Por que essa diferença?
  - Várias possíveis causas foram investigadas
    - Número de pacientes, práticas religiosas, variações climáticas,...
    - Única diferença clara era o propósito de cada clínica



## Mais recentemente

- Papel de cada clínica
  - Primeira clínica:
    - Formação de estudantes de medicina
  - Segunda clínica:
    - Formação de parteiras

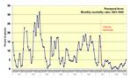


## Parto e dissecação



## Mais recentemente

- Em 1847 um colega de Semmelweis faleceu com uma patologia semelhante
  - Após ferimento durante uma autópsia com o instrumento utilizado na dissecação
  - Semmelweis associou contaminação por cadáver à febre puerperal
    - Pediu que os estudantes lavassem as mãos com água clorada entre autópsia e exame de paciente
      - Taxa de mortalidade na primeira clínica (que tinha a maior taxa) caiu 90%



## Resultado

- Proposta de Semmelweis foi rejeitada por outros obstetras
  - Por causa de seu temperamento difícil e de críticas de colegas, foi internado em um asilo
    - Faleceu 13 dias depois
    - Infecção em um ferimento na mão
      - Mal semelhante ao que ajudou a evitar



## Resultado

- Semmelweis é hoje chamado de salvador das mães
- Universidade de Medicina de Budapeste
  - Primeira faculdade de medicina da Hungria foi criada em 1769
  - Em 1969 passou a se chamar Semmelweis University



## Melhorou?

- Século XIX
  - Hospital mais seguro do Reino Unido: 4 mães morriam a cada 100 partos
    - 18 em Viena (não lavavam as mãos)
- Hoje
  - Reino Unido: em média, 1 mãe morre a cada 12.000 partos
    - Brasil: em média, 1 mãe morre a cada 1.500 partos



## Melhorou?

- Século XIX
  - Hospital mais seguro do Reino Unido: 4 mães morriam a cada 100 partos
    - 18 em Viena (não lavavam as mãos)
- Hoje
  - Reino Unido: em média, 1 mãe morre a cada 12.000 partos
    - Brasil: em média, 1 mãe morre a cada 1.500 partos
  - Mundo: < 30% dos médicos lavam as mãos

## Análise de dados

- Como era:
  - Grande parte das teorias eram validadas em quantidades muito pequenas de dados
  - Previsões feitas por computadores tinham baixa taxa de acerto
    - Em 1971, foi afirmado que terremotos poderiam ser previstos 10 anos antes
    - Em 1989, foi previsto que aquecimento global iria aumentar em 2o Celsius a temperatura média anual dos EUA



## Análise de dados

- A previsão por seres humanos é ainda pior
  - Os americanos precisam de telefones, nós não, pois temos muitos mensageiros (1878)
    - William Preece, Engenheiro chefe dos correios britânicos
  - É impossível que o iPhone tenha mercado (2007)
    - Steve Ballmer, presidente da Microsoft
  - Não existem tantos filmes assim que eu queira assistir (2005)
    - Steve Chen, co-fundador do YouTube

## Poucos dados

- Difícil extrair modelos confiáveis
- Notícia verdadeira de jornal:
  - 50% dos dentistas de uma cidade morrem em um acidente de ônibus



## Poucos dados

- Difícil extrair modelos confiáveis
- Notícia verdadeira de jornal:
  - 50% dos dentistas de uma cidade morrem em um acidente de ônibus
    - A cidade tinha 2 dentistas



## Análise de dados

- Análise dos dados por seres humanos
  - Falta de especialistas
  - Custo elevado
  - Subjetividade
  - Dificuldade de lidar com grande volume
- Técnicas tradicionais para análise
  - Planilhas
  - Sistemas de gerenciamento de bancos de dados



## Análise 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?
    - Qual a estrutura terciária de uma nova proteína
  - Técnicas mais sofisticadas podem extrair conhecimento de grandes conjuntos de dados



## Análise de Dados

- Investigada pela Ciência de Dados
- O que é a ciência de dados?
  - Várias definições
    - Estuda princípios, métodos e sistemas computacionais para extrair conhecimento de dados
  - Pergunta chave da área:
    - Como encontrar de forma eficiente padrões em (grandes) conjuntos (fluxos) de dados



## Ciência de Dados

- Teorias e princípios gerais ainda estão sendo formulados
- Área basicamente experimental
- Mas a mudança está sendo rápida
  - Livro recente propõe nova forma de abordar teoria da computação
    - Baseada em dados

## Big Data x Ciência de Dados

- Os termos Big Data e Ciência de Dados são frequentemente confundidos
  - Confusão ocorre principalmente por interesses mercadológicos
  - Ciência de Dados procura criar modelos capazes de extrair padrões de sistemas complexos
    - E usar esses modelos em aplicações reais
  - Big Data procura dar suporte à coleta e ao gerenciamento de grandes quantidades de dados

**Colecionar x Descobrir**



## Big Data x Ciência de Dados

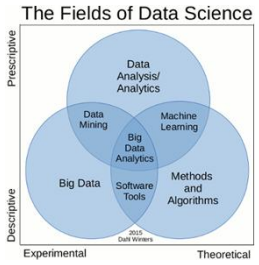
- Big data
  - Ferramentas e ambientes computacionais
  - Inclui bancos de dados e sistemas distribuídos
- Ciência de dados
  - Dá suporte à tomada de decisão orientada por dados
    - *Data-driven decision making* (DDD)
      - Baseia decisões na análise de dados
        - Ao invés de apenas na experiência ou intuição
        - Duas formas podem ser combinadas



## Ciência de Dados




## Ciência de dados






## Perigo?

http://www.time.com/time/magazine/article/0,9171,2058205,00.html



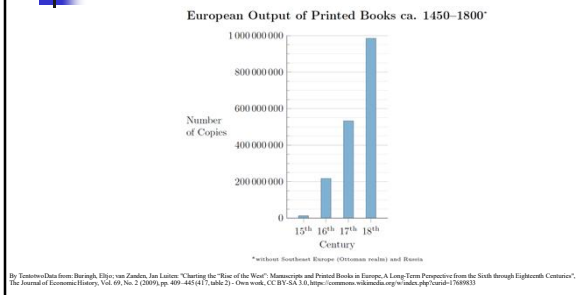
## Perigo?

- Quando o disponibilidade de dados supera a capacidade de processá-los
  - Máquina de impressão de Gutenberg
    - Reforma de Martinho Lutero causou várias guerras

**Problema tecnológico ou humano ?**

Jost Amman - Meggs, Philip B. *A History of Graphic Design*. John Wiley & Sons, Inc. 1998. (p 64)

## Perigo?




By Tentotwo;Data from: Buringh, Eltjo; van Zanden, Jan Luiten: "Charting the "Rise of the West": Manuscripts and Printed Books in Europe, A Long-Term Perspective from the Sixth through Eighteenth Centuries", The Journal of Economic History, Vol. 69, No. 2 (2009), pp. 409–445 (417, table 2) - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=17688933



## Vídeos



## Riscos e benefícios

- Toda invenção, incluindo roda e fogo, pode trazer riscos e benefícios
  - Reduzir exposição a situações de perigo
  - Reduzir a realização de tarefas repetitivas, monótonas e arriscadas
  - Melhorar a produtividade, permitindo fazer mais e melhor com menos esforço
  - Reduzir custos, permitindo que mais pessoas tenham acesso a um serviço ou produto
    - Saúde, educação, moradia, saneamento



## Ciência de Dados

- Mineração de dados é o conceito mais semelhante a ciência de dados
  - Mas CD é muito mais que isso
- Inclui
  - Planejamento de experimentos
  - Pré-processamento
  - Modelagem
  - Avaliação

Pré-processamento, Modelagem: Mineração de dados



## Ciência de Dados

- Desafios para extração do significado real dos dados

Dados ruins | Falhas no processo | Interpretação incorreta dos resultados

→ Falsos positivos e negativos

→ Conclusões equivocadas

→ Decisões erradas



## Qualidade de dados

- Em geral, dados não foram gerados para uso em Ciência de Dados
  - Produzidos para outros propósitos
  - Frequentemente apresentam problemas
- Algoritmos de modelagem precisam geralmente de dados "limpos"
  - Entra lixo, sai lixo
  - Problemas nos dados precisam ser detectados e corrigidos
    - Limpeza de dados

## Modelagem

- Indução
  - Processo de extrair uma hipótese (ou teoria) a partir de observações
- Dedução
  - Processo de raciocínio em que a partir de conhecimentos gerais, um conhecimento geral é obtido a partir de conhecimentos específicos



## Qualidade dos modelos

- Depende das suposições feitas para seu desenvolvimento
  - Reator nuclear de Fukushima
    - Projetado para lidar com terremoto de magnitude de até 8,6 na escala Richter
      - Terremoto em 03/2011 teve magnitude de 8.9
  - Ciclovia do Rio de Janeiro
    - Projetada para ondas de 2,5 metros
      - Onda em 04/2016 atingiu 50 metros



## Replicação é essencial

- Estudo da Bayer
- Mais de 2/3 das descobertas científicas publicadas em periódicos são falsas
  - Não puderam ser replicadas



## Ciência de Dados para o Bem

- Movimentos sem fins lucrativos
  - Para trazer benefícios sociais para pessoas e comunidades carentes
  - Alguns deles são adotados pelas empresas
- Como isso ocorre?
  - Reuniões
  - Eventos
  - Estágios acadêmicos
  - Redes sociais



## Ciência de Dados para o Bem

- Abordagens existentes:
  - Uso de dados (abertos) para resolver problemas de defesa civil
    - Normalmente, desenvolvimento de aplicativos móveis / web
  - Uso de ciência de dados para resolver problemas sociais
    - Principalmente buscando insights de cientistas de dados



## Ciência de Dados para o Bem

- Abordagens existentes:
  - Democratização de dados
    - Permitir que qualquer pessoa tenha acesso a dados
    - Primeiro Cientista Chefe de Dados foi nomeado em 2015 pelo presidente dos EUA
      - First U.S. Chief Data Scientist
      - Estimular pesquisas e desenvolvimento tecnológico em medicina de precisão, dados abertos e decisão apoiada por dados

## Ciência de Dados para o Bem

- Diferentes formas de engajamento
  - Desafios e competições
    - Análise de dados preditivos para prevenção de incêndios
    - http://ibmhadoop.devpost.com/
  - Estágios universitários
  - Trabalho voluntário
  - Trabalho de meio período
  - Empregos de turno completo


http://www.kdnuggets.com/2014/07/data-for-good-data-driven-projects-social-good.html



## Ciência de Dados para o Bem

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



## Ciência de Dados para o Bem

- Educação
  - Monitorar o desempenho dos alunos
    - Apoiar o desenvolvimento de melhores plataformas de ensino
    - Ensino dinamicamente adaptado para o desempenho e as necessidades estudantes
  - Avaliar professores e escolas
    - Replicar boas experiências
    - Agir antes que seja tarde



## Ciência de Dados para o Bem

- Finanças
  - Melhorar a situação financeira de comunidades carentes
  - Apoiar pequenas empresas e cooperativas
  - Direcionar iniciativas sociais
  - Detecção de fraude no uso de recursos públicos



## Ciência de Dados para o Bem

- Meio ambiente
  - Reduzir índices de poluição
  - Diminuir o desmatamento
  - Reduzir os efeitos de secas e enchentes
  - Prever ocorrência de catástrofes
  - Detectar espécies invasoras
  - Aumentar a diversidade de espécies



## Ciência de Dados para o Bem

- Saúde
  - Monitorar o estado do paciente em unidades de terapia intensiva
  - Acelerar avanços e tornar a pesquisa médica mais barata
  - Analisar milhões de registros de pacientes que chegam em fluxos de dados
  - Identificar a ocorrência de epidemias
  - Prevenção de quedas de idosos

## Ciência de Dados para o Bem

- Vídeo

https://www.youtube.com/watch?v=V_ndznCKjg4



## Ciência de Dados para o Bem

- Links relevantes
  - Data Science for Social Good Fellowship
  - DataLook
  - civisanalytics.com
  - digitalhumanitarians.com
  - www.data4good.co
  - http://www.meetup.com/DataKind-UK



## Conclusão

- Por que todo esse interesse?
  - Cada vez mais dados são gerados
  - Conhecimento precioso (vantagem competitiva) presente nesses dados
  - Avanços na capacidade de processamento e pervasividade de 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
- Mais arte que ciência
- Baseada em
  - Computação
    - Aprendizado de Máquina
  - Estatística



## Conclusão

- Análise de dados
- Ciência de dados
- Big data
- Crescimento da área
- Ciência de dados para o bem



## Exercício

- Descrever o conjunto de dados a ser usado no curso

# Perguntas