UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Prof. Heitor Silvério Lopes
Prof. Thiago H. Silva

Mineração de Dados e Descoberta de Conhecimento

Aula 1a: Introdução (2023)

Mineração de Dados (CPGEI & PPGCA)

- **Professores**: **Heitor S. Lopes e Thiago H. Silva**
- **Carga Horária**: ~45 horas-aula
- **Horário**: 3ª feira, ~08:20-12:00
- **Ementa**: Motivações. O processo de descoberta do conhecimento a partir de dados. Tipos de dados e sua análise. Preparação, limpeza e visualização de dados. Estudo de algoritmos para as principais tarefas de mineração de dados: classificação, associação, agrupamento, regressão linear, detecção de anomalias, seleção de atributos, redução de dimensionalidade. Mineração de textos, imagens, sinais e séries temporais
- **Avaliações**: Exercícios: 30%, Projeto final da disciplina: **70%**, constituído de desenvolvimento de projeto de mineração de dados, elaboração de relatório escrito e apresentação em seminário
- **Bibliografia básica**:
- P-N. Tan; M. Steinbach; V. Kumar. Introduction to Data Mining. 2nd edition. Pearson, 2018.

- **Sites da disciplina**:

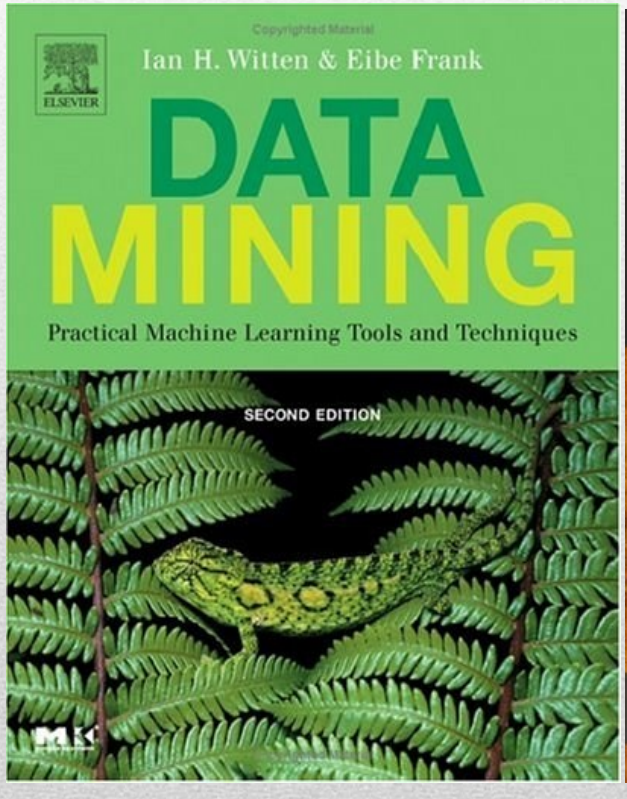
  Moodle: https://moodle.utfpr.edu.br/course/view.php?id=2702

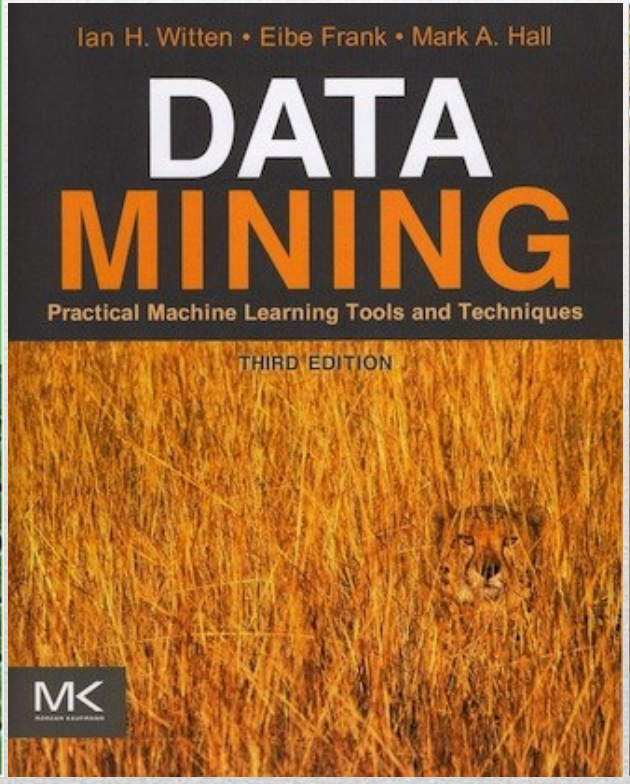
  Site: http://silverio.net.br/heitor/disciplinas/md

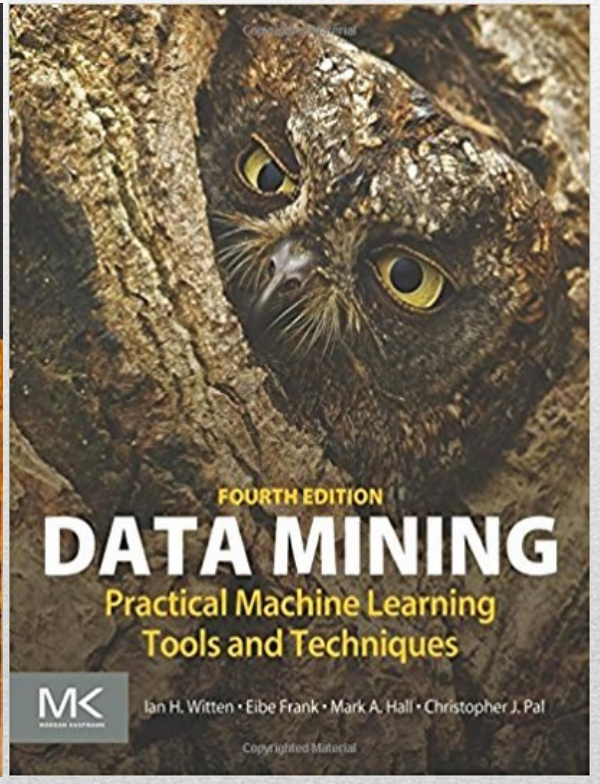
DATA MINING
know it all
Jiawei Han
Micheline Kamber
Data Mining
Concepts and Techniques
INTRODUCTION TO DATA MINING
PANG-NING TAN
MICHAEL STEINBACH
VIPIN KUMAR
Second Edition
Data Mining
Concepts and Techniques
Jiawei Han and Micheline Kamber
DATA MINING ALGORITHMS
Narosa
Rajan Chattamvelli
DATA MINING FOR INTELLIGENCE, FRAUD, & CRIMINAL DETECTION
Advanced Analytics & Information Sharing Technologies
CHRISTOPHER WESTPHAL
WILEY
Data Mining Cookbook
Modeling Data for Marketing, Risk, and Customer Relationship Management
Olivia Parr Rud
Knowledge Discovery from Data Streams
Joao Gama
Data Mining and Data Warehousing
PRINCIPLES OF
Data Mining
David Hand
Heikki Mannila
Padhraic Smyth
Data Mining for Bioinformatics
Sumeet Dua
Pradeep Chowriappa
CRC Press
DATA MINING FOR GENOMICS AND PROTEOMICS
ANALYSIS OF GENE AND PROTEIN EXPRESSION DATA
DARIUS M. DZIUDA
WILEY
Jason T. L. Wang
Mohammed J. Zaki
Hannu T. T. Toivonen
Dennis Shasha (Eds)
Data Mining in Bioinformatics
Springer
Chapman & Hall/CRC
Data Mining and Knowledge Discovery Series
Biological Data Mining
Edited by
Jake Y. Chen
Stefano Lonardi
CRC Press
JAVA DATA MINING
Strategy, Standard, and Practice
A Practical Guide for Architecture, Design, and Implementation
Mark F. Hornick
Erik Marcadé
Sunil Venkayala

Mineração de dados
→ Descoberta de conhecimento

- Encontrar conhecimento novo, útil e relevante, "oculto" entre dados

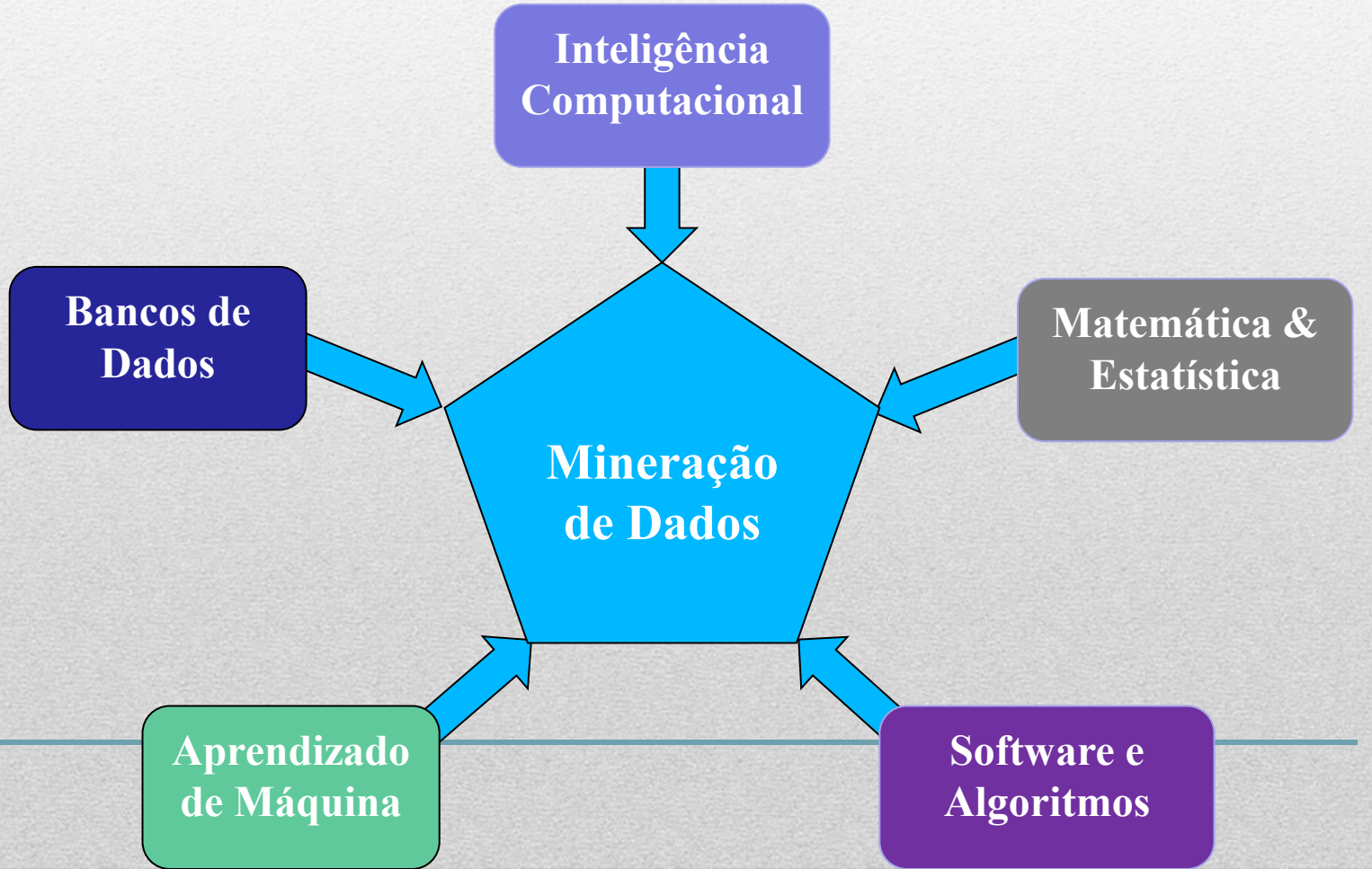






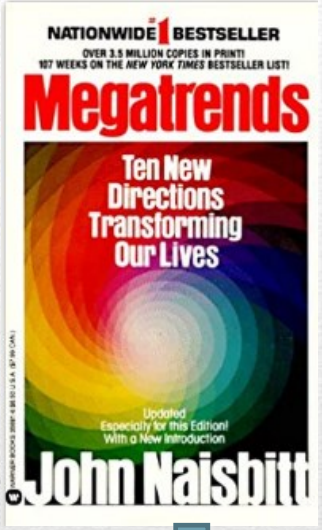
Multidisciplinaridade da Mineração de Dados

- ✓ A Mineração de Dados utiliza conceitos e técnicas de diversas áreas:



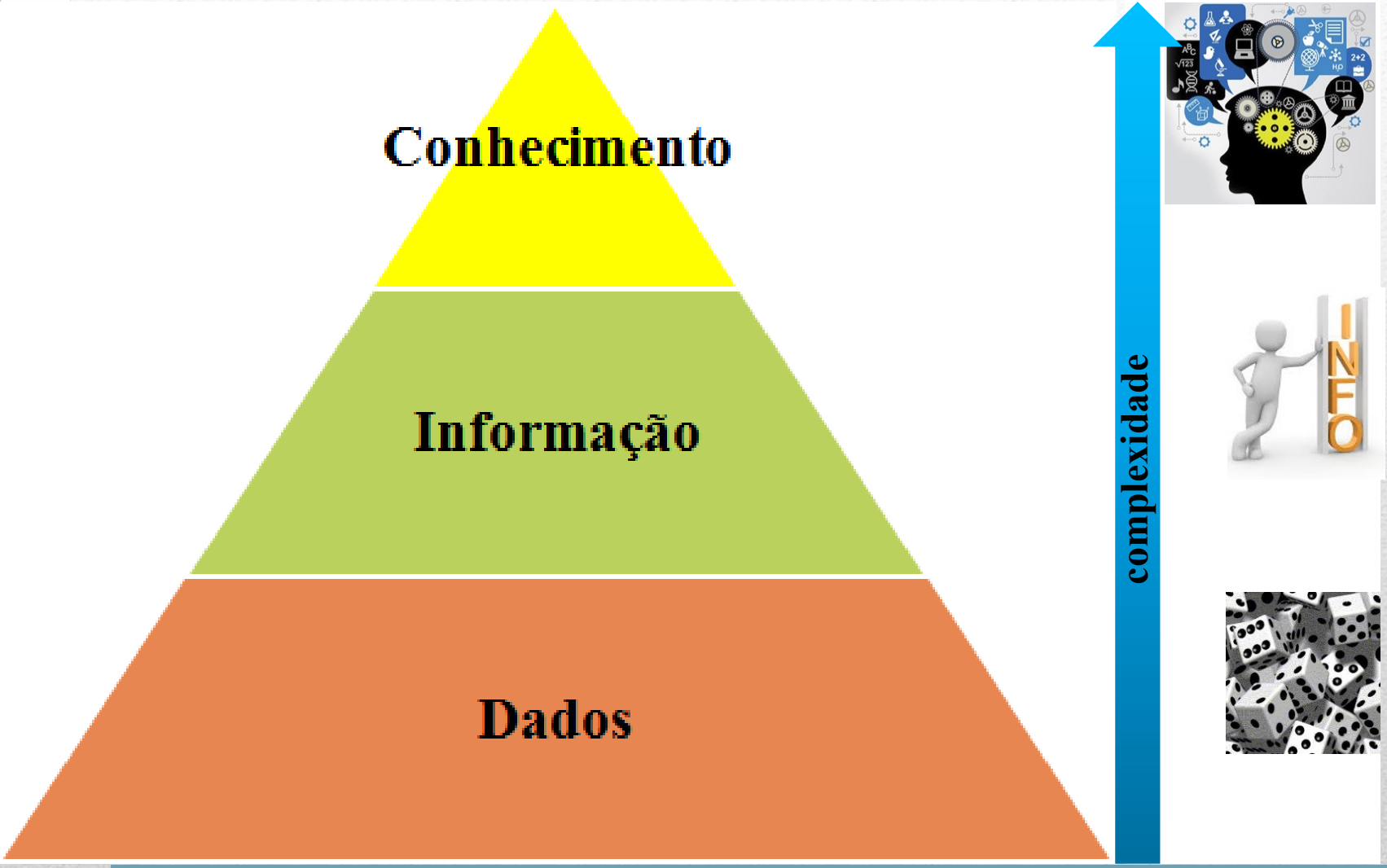
Dados X Informação X Conhecimento

- ✓ **Dados:**
  - Instâncias (objetos, pessoas, instantes de tempo, etc)
  - Descrevem propriedades individuais e não coletivas, e são:
    - Disponíveis em grande quantidade e múltiplas formas
    - Fáceis de coletar
    - Pouco úteis para fazer predições ou tomar decisões
- ✓ **Informação:**
  - Se refere a classes de instâncias (grupos de objetos, pessoas,
  - Descreve padrões genéricos, estruturas, leis, princípios, etc
  - Usualmente são:
    - Pouco abundantes
    - Difícil de obter
    - Permite generalizações e predições
- ✓ **Conhecimento:**
  - É a compreensão de alguma coisa (incluindo fatos, habilidades e informações), obtido por percepção, descoberta ou aprendizado



We are drowning in information, but starving for knowledge.
John Naisbitt (**1982**)

Dados X Informação X Conhecimento



Definições de Mineração de Dados

1. (Witten et al., 2000)
   - Descoberta automática/semi-automática de padrões estruturais em dados
2. (Fayyad et al., 1996)
   - Extração de conhecimento estruturado útil, previamente desconhecido, não-trivial, compreensível e implícito a grandes massas de dados.

✓ Critérios desejáveis para o conhecimento descoberto:

- **Corretude, Generalidade, Utilidade, Compreensibilidade, Novidade**

Mineração de Dados e Descoberta de Conhecimento

- ✓ Seja um dataset dos prontuários de pacientes de uma maternidade. Um procedimento de mineração de dados pode chegar à seguinte regra:
- ✓ **SE** (paciente.idade>15 **E** paciente.idade<50) **E** (atendimento.setor="clínica cirúrgica") **E** (cirurgia.tipo="cesariana") **ENTÃO** paciente.sexo="Feminino"

Corretude ☺
Generalidade ☺
Utilidade ☹
Compreensibilidade ☺
Novidade ☹

- ✓ Seja um dataset dos prontuários de oncológicos infantis *. Um procedimento de mineração de dados pode chegar à seguinte regra:
- ✓ **SE** (histologia=carcinoma) **E** paciente.idade<3) **E** (estágio=1) **E** (metástase="não") **ENTÃO** anos.sobrevida>5

Corretude ☺
Generalidade ☺
Utilidade ☺ ☺
Compreensibilidade ☺
Novidade ☺ ☺

* Bojarczuk, C.C., Lopes, H.S., Freitas, A.A. A constrained-syntax genetic programming system for discovering classification rules: application to medical data sets. **Artificial Intelligence in Medicine**, v. 30, n. 1, p. 27-48, 2004.

Ciclo de vida de projetos de Mineração de dados

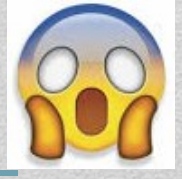
Dados brutos

Coleta, seleção e integração de dados

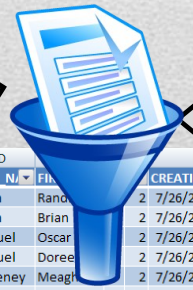
Data warehouse

Pré-processamento: formatação, limpeza, redução dos dados

>60% do esforço

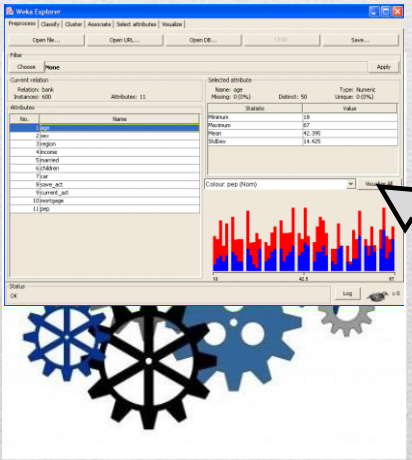


Dados filtrados

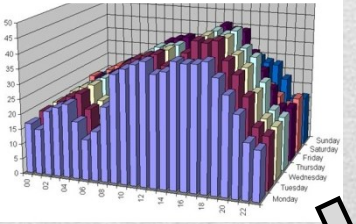


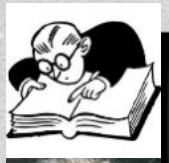
	A	B	C	D			H	I	J
1	CO	ME	CON	LAST_NA	FIN		CREATION_D	I	LAST_UPDAT
2	1001	1001	1003	Winn	Rand	2	7/26/2008 7:23	2	7/26/2008 7:23
3	1002	1001	1003	Winn	Brian	2	7/26/2008 7:23	2	7/26/2008 7:23
4	1003	1002	1003	Vizquel	Oscar	2	7/26/2008 7:23	2	7/26/2008 7:23
5	1004	1002	1003	Vizquel	Doree	2	7/26/2008 7:23	2	7/26/2008 7:23
6	1005	1003	1003	Sweeney	Meagh	2	7/26/2008 7:23	2	7/26/2008 7:23
7	1006	1003	1003	Sweeney	Matthew	2	7/26/2008 7:23	2	7/26/2008 7:23
8	1007	1003	1003	Sweeney	Ian M	2	7/26/2008 7:23	2	7/26/2008 7:23

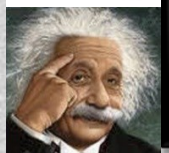
Mineração de dados



Padrões





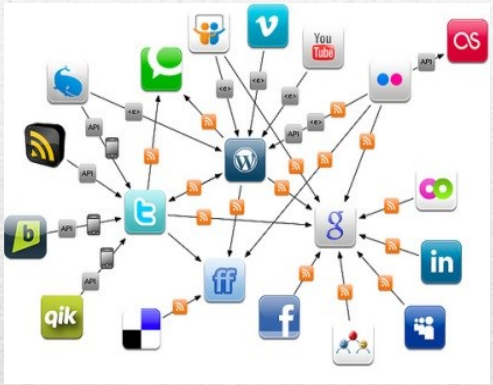


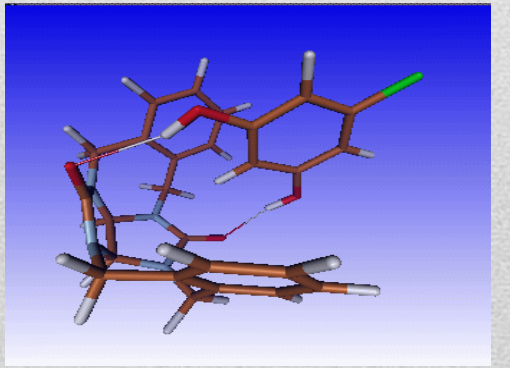
Análise e interpretação dos padrões

Motivações - Mineração de Dados (1)

✓ <u>Grande quantidade</u> de dados gerada:

- E-mails e redes sociais
- Transações comerciais e bancárias
- Buscas em páginas web
- Dados médicos e biológicos
- Dados científicos e astronômicos





Qual o “tamanho” da internet ?

4,938,743,302	1,863,954,456	163,420,834,022
Internet Users in the world	Total number of Websites	Emails sent today
4,795,887,808	4,629,219	492,476,084
Google searches today	Blog posts written today	Tweets sent today
4,661,081,986	55,784,382	98,307,303

(http://www.internetlivestats.com/)

04/06/21 14:24

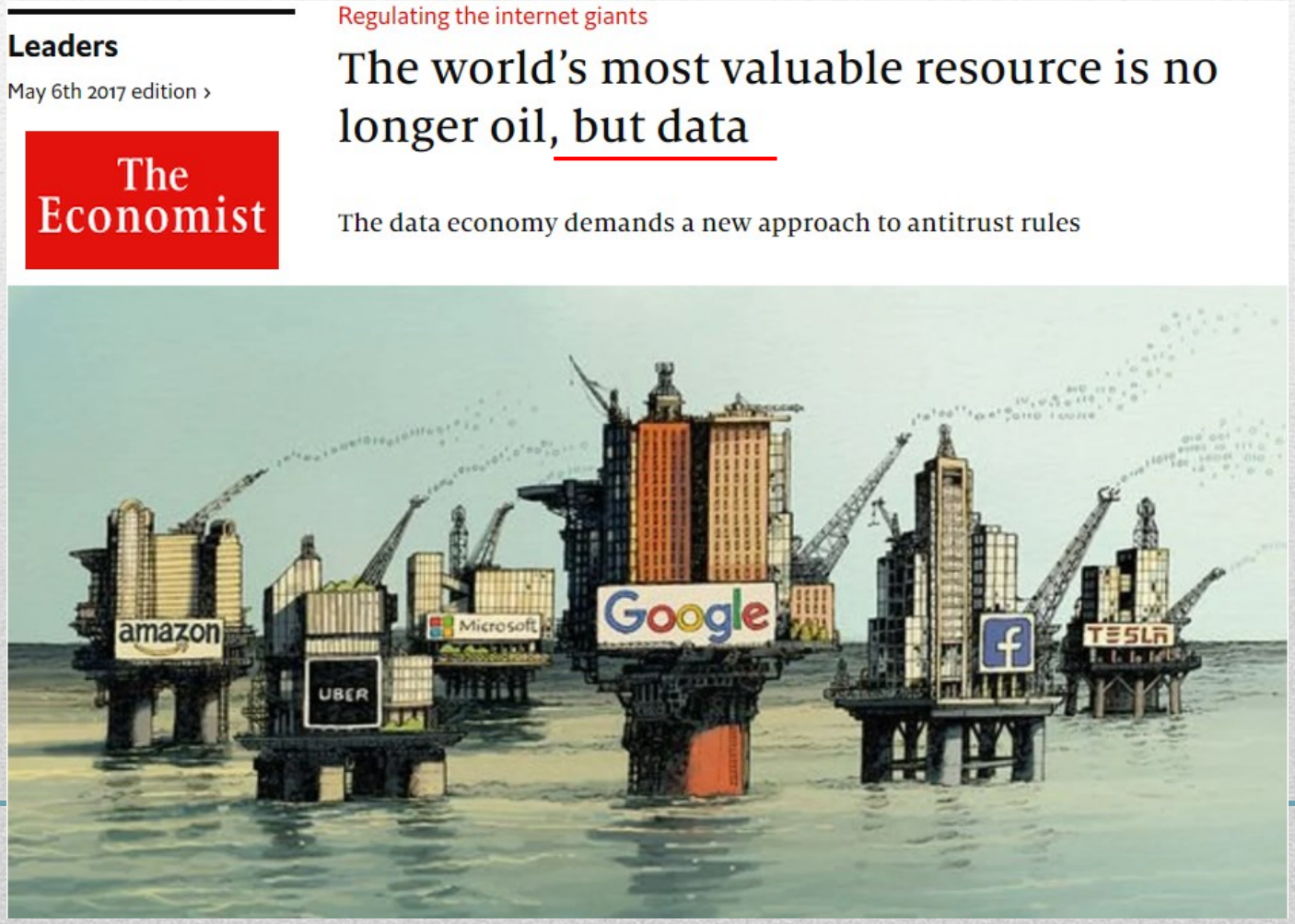
Qual o “tamanho” da internet ?

5,322,318,568	1,955,741,104	267,196,458,460
Internet Users in the world	Total number of Websites	Emails sent today
8,375,132,140	8,146,699	817,788,980
Google searches today	Blog posts written today	Tweets sent today
7,815,459,960	94,892,250	171,516,988

31/05/22 22:40

Motivações - Mineração de Dados (2)

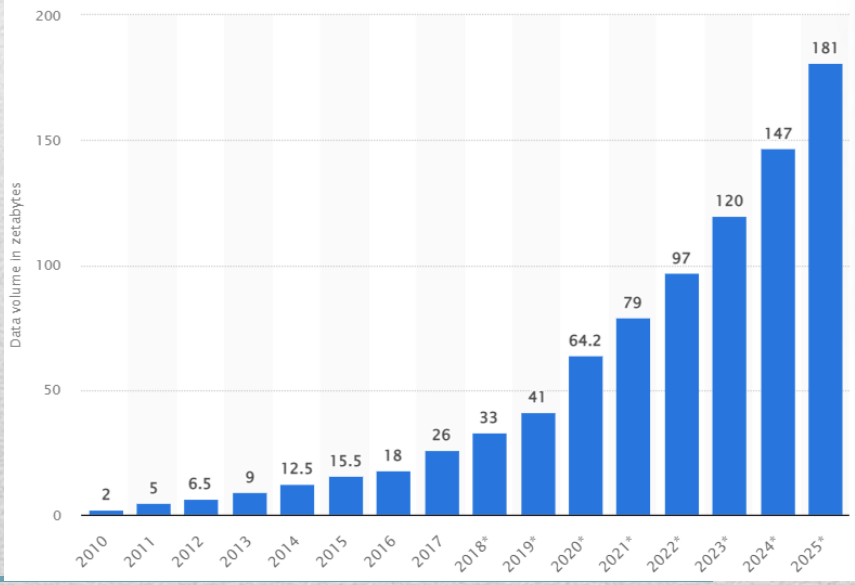
✓Maior motivação: Interesse comercial ($$$$$$$$)



Dilema crítico: Quantidade x Utilidade (de dados)

- ✓ O volume de dados gerado, criado, capturado,... cresce **exponencialmente**
- ✓ A capacidade para entender e efetivamente utilizar estes dados cresce **linearmente**

A mineração de dados pode nos ajudar a <u>compreender</u> grandes massas de dados e <u>extrair</u> algum conhecimento útil



Algumas aplicações de Mineração de Dados

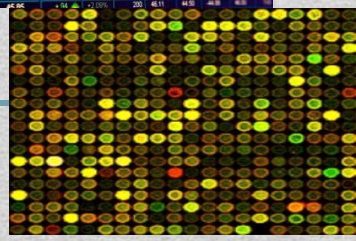
- ✓ Descobrir os interesses de consumo dos clientes para aumentar as vendas
- ✓ Compreender o funcionamento dos diversos sistemas do corpo humano e projetar novos medicamentos

- ✓ Descobrir como o comportamento dos eleitores é influenciado

- ✓ Descobrir irregularidades nos gastos públicos

- ✓ Prever o comportamento do mercado de capitais e da economia

- ✓ Descobrir novos conhecimentos !



Questões éticas em MD & KDD

IMPORTANTE

- ✓ Questões importantes relacionadas aos dados:
  - A quem pertencem os dados ?
  - Quem pode permitir acesso aos dados?
  - Quem pode acessar os dados?
  - Para qual finalidade os dados foram coletados?
  - Há informações sensíveis nos dados ?
    - **Anonimização** é importante para não identificar pessoas/empresas/instituições, etc

- ✓ Métodos de MD&KDD podem gerar resultados contestáveis quando baseados e pequenas amostras ou métodos inadequados. Qual o “tamanho” ideal ??
- ✓ MD&KDD não deve ser utilizado para discriminar pessoas com base na classe social, sexo, religião, etc

MAPA DA SAÚDE

Google vai seguir 10 mil pessoas para estudar doenças

DA AFP - A Verily, a filial especializada em saúde da Alphabet, empresa-mãe do Google, vai acompanhar 10 mil pessoas durante anos para detectar sinais que possam antecipar o surgimento de doenças.

O estudo, chamado "Project Baseline", será realizado em colaboração com as universidades de Duke e Stanford. O objetivo é criar uma espécie de mapa da saúde humana por meio da identificação de uma ampla gama de dados médicos, genéticos e de comportamento.

Os participantes usarão diariamente um dispositivo conectado no pulso e outros sensores. O acompanhamento será feito com visitas a clínicas e pesquisas via smartphones.

A Verily (antiga Google Life Science) disse também que recolherá amostras biológicas (sangue, saliva), dados clínicos, e genéticos e imagens médicas.

O projeto espera registrar os marcadores que indicam que uma doença cardíaca ou um câncer está se desenvolvendo.

A informação será guardada na "nuvem" do Google, e a Verily pretende colocar as bases de dados anônimos à disposição dos pesquisadores.

Existe privacidade de dados ???

Coisas que você cria

Se você está conectado à sua Conta do Google, armazenamos e protegemos aquilo que você criar usando nossos serviços. Isso pode incluir:

- E-mails enviados e recebidos no Gmail
- Contatos que você adiciona
- Eventos da agenda
- Fotos e vídeos dos quais você faz o upload
- Documentos, Planilhas e Apresentações no Drive

Coisas que você faz

Quando você acessa nossos serviços (por exemplo, faz uma pesquisa no Google, recebe rotas no Google Maps ou assiste um vídeo no YouTube), coletamos dados para fazer com que esses serviços funcionem melhor para você. Isso pode incluir:

- Coisas que você pesquisa
- Sites que você visita
- Vídeos a que você assiste
- Anúncios nos quais você clica ou toca
- Seu local
- Informações do dispositivo
- Endereço IP e dados de cookies

Coisas que fazem parte de você

Quando você se inscreve em uma Conta do Google, nós armazenamos as informações básicas que você nos fornece. Isso pode incluir:

- Nome
- Endereço e senha de e-mail
- Data de nascimento
- Gênero
- Número de telefone
- País

Brave

Tor

Dados abertos (ver mais na página da disciplina)

- ✓ Iniciativa mundial que permite livre acesso a dados colocados em domínio público
- ✓ Dados abertos de Curitiba (http://www.curitiba.pr.gov.br/dadosabertos/)
- ✓ Dados abertos de São Paulo (http://dados.prefeitura.sp.gov.br/)
- ✓ Dados abertos do Rio de Janeiro (http://data.rio/)
- ✓ Portal Brasileiro de Dados Abertos (http://dados.gov.br/)

- ✓ The U.S. Government's open data (https://www.data.gov/)
- ✓ Dados científicos (https://www.pangaea.de/)
- ✓ NASA open data (https://open.nasa.gov/open-data/)
- ✓ Digital sky survey (http://www.sdss.org/)

Tarefas e Métodos Importantes em Mineração de dados

Tarefas	Métodos
Classificação	Árvores de decisão (C4.5, CN2), regras de classificação, k-NN, Random Forest, SVM, classificadores Bayesianos, redes neurais, Adaboost
Análise de dependências	Apriori, FP-growth, Eclat, zigzag
Regressão	Regressão linear, regressão polinomial, regressão logística
Redução de dimensionalidade & seleção de atributos	Principal Component Analysis, Chi-quadrado, Information Gain, Symmetric Uncertainty, Manifold learning
Visualização de dados	Gráfico de silhueta, scatter plot, heatmap, box plot, clusters, t-SNE
Agrupamento (clusterização)	K-means, Mapa auto-organizado de Kohonen (SOM), DBScan, agrupamento hierárquico
Sinais e séries temporais	ARMA, ARIMA, ARIMAX
Mineração de textos	Bag-of-words, word-to-vector, análise de sentimentos

Principais softwares de código aberto para Mineração de Dados

- **Weka** (Java): Waikato University (NZ) https://www.cs.waikato.ac.nz/ml/weka

- **Orange** {Python): University of Ljubljana (Slovenia) https://orangedatamining.com/

- **Knime** (Java): Konztanz Universitaet https://www.knime.com/