Introdução à Mineração de Dados com Aplicações em Ciências Espaciais

Escola de Verão do Laboratório Associado de Computação e Matemática Aplicada

Rafael Santos

Programa

- ***Dia 1:*** Apresentação dos conceitos de mineração de dados, motivação e alguns exemplos.
- ***Dia 2:*** Algoritmos de classificação supervisionada e aplicações.
- ***Dia 3:*** Algoritmos de classificação não-supervisionada e aplicações. Algoritmos de mineração de associações.
- ***Dia 4:*** Visualização e mineração de dados. Outros algoritmos e idéias. Onde aprender mais.

Objetivo

- Apresentar conceitos, técnicas e exemplos de aplicação de mineração de dados.
- Descrever alguns dos algoritmos mais utilizados com exemplos de aplicação.
- Parte reduzida do material da disciplina CAP-359 do Programa de Pós-Graduação em Computação Aplicada.
- ***Math-Lite!***

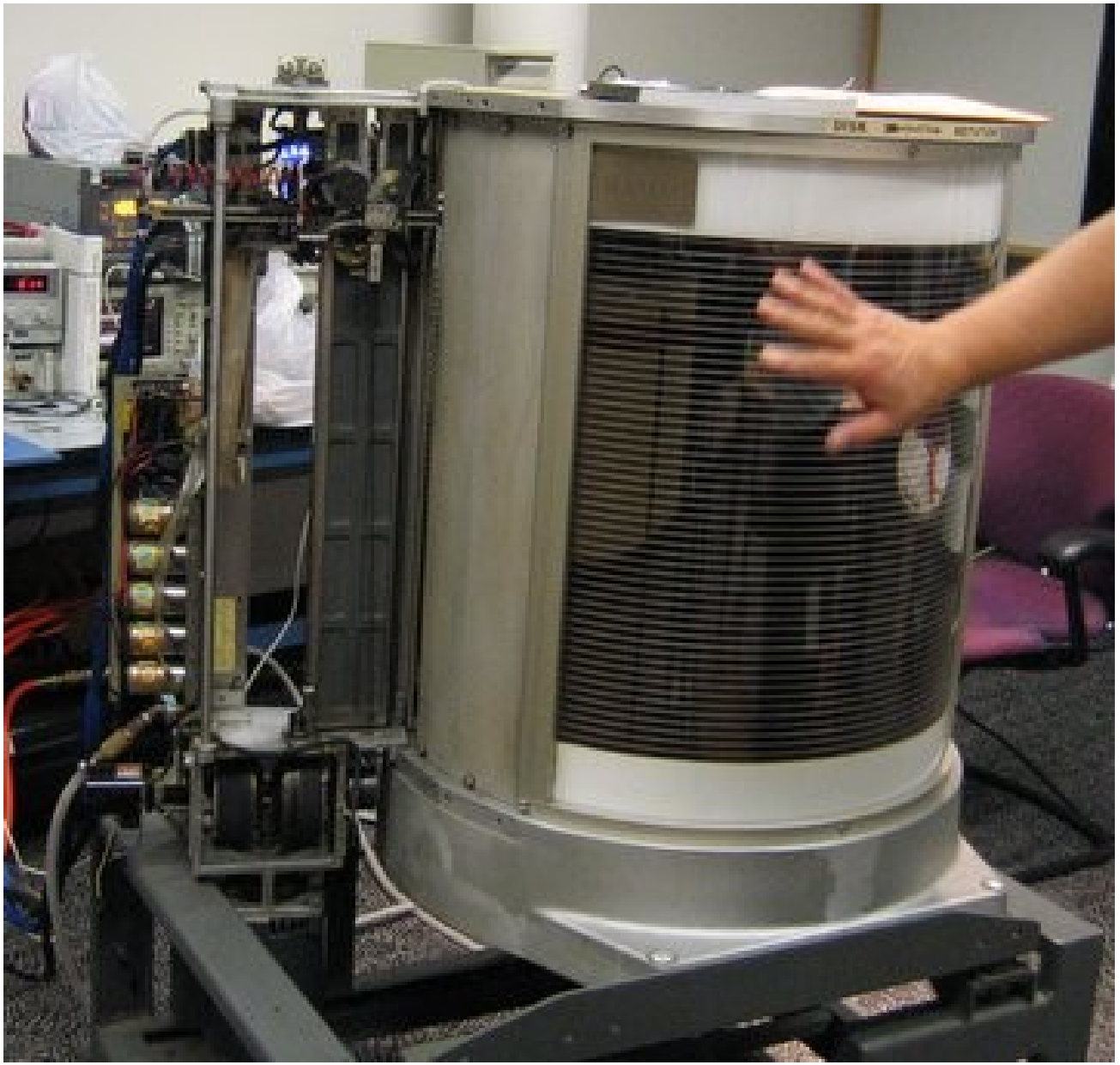
Introdução e Motivação

O *Tsunami* de Dados

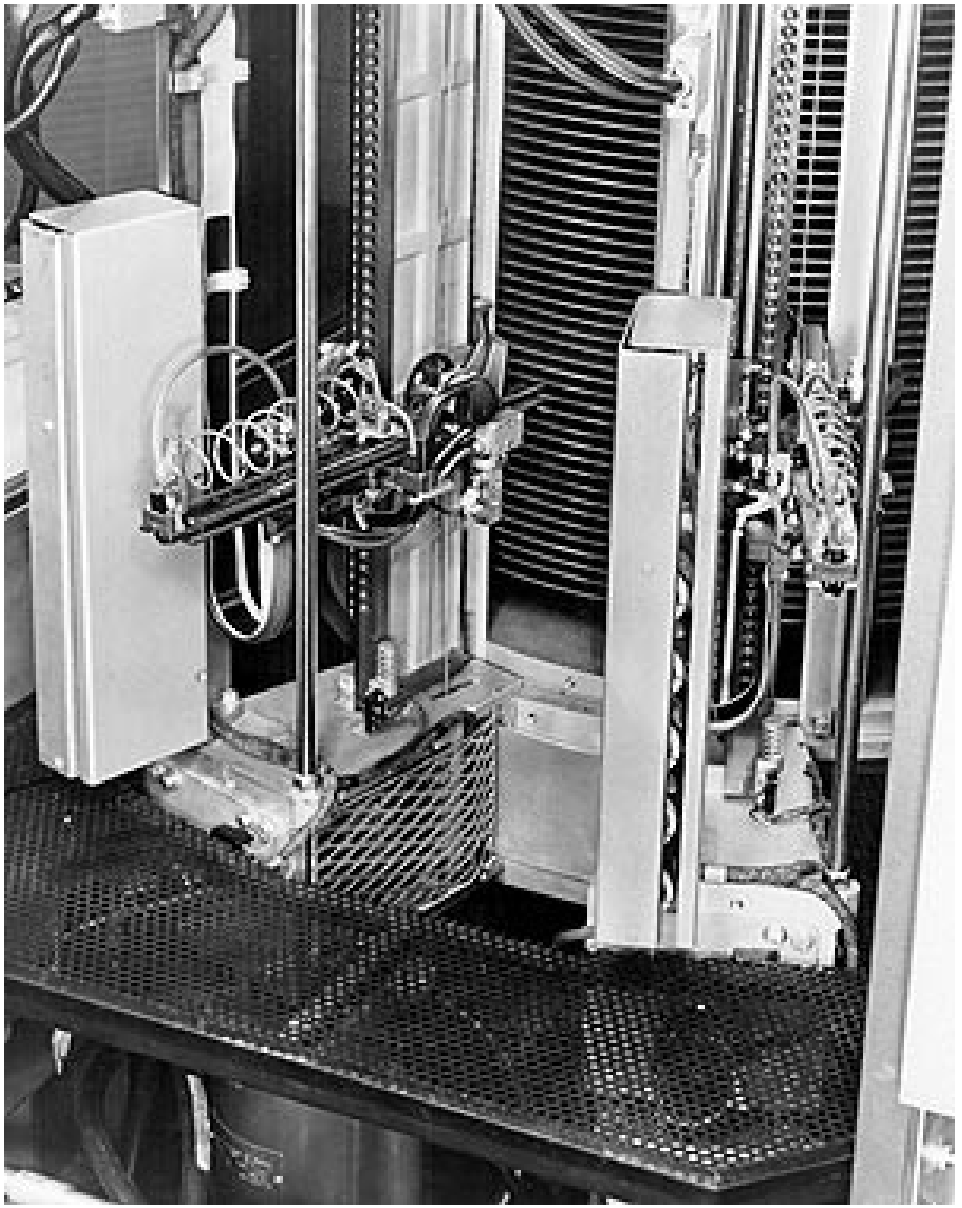
"We are drowning in information but starved for knowledge." –
John Naisbitt, Megatrends (1984).

O que é e como nos afeta?

Introdução e Motivação



Introdução e Motivação

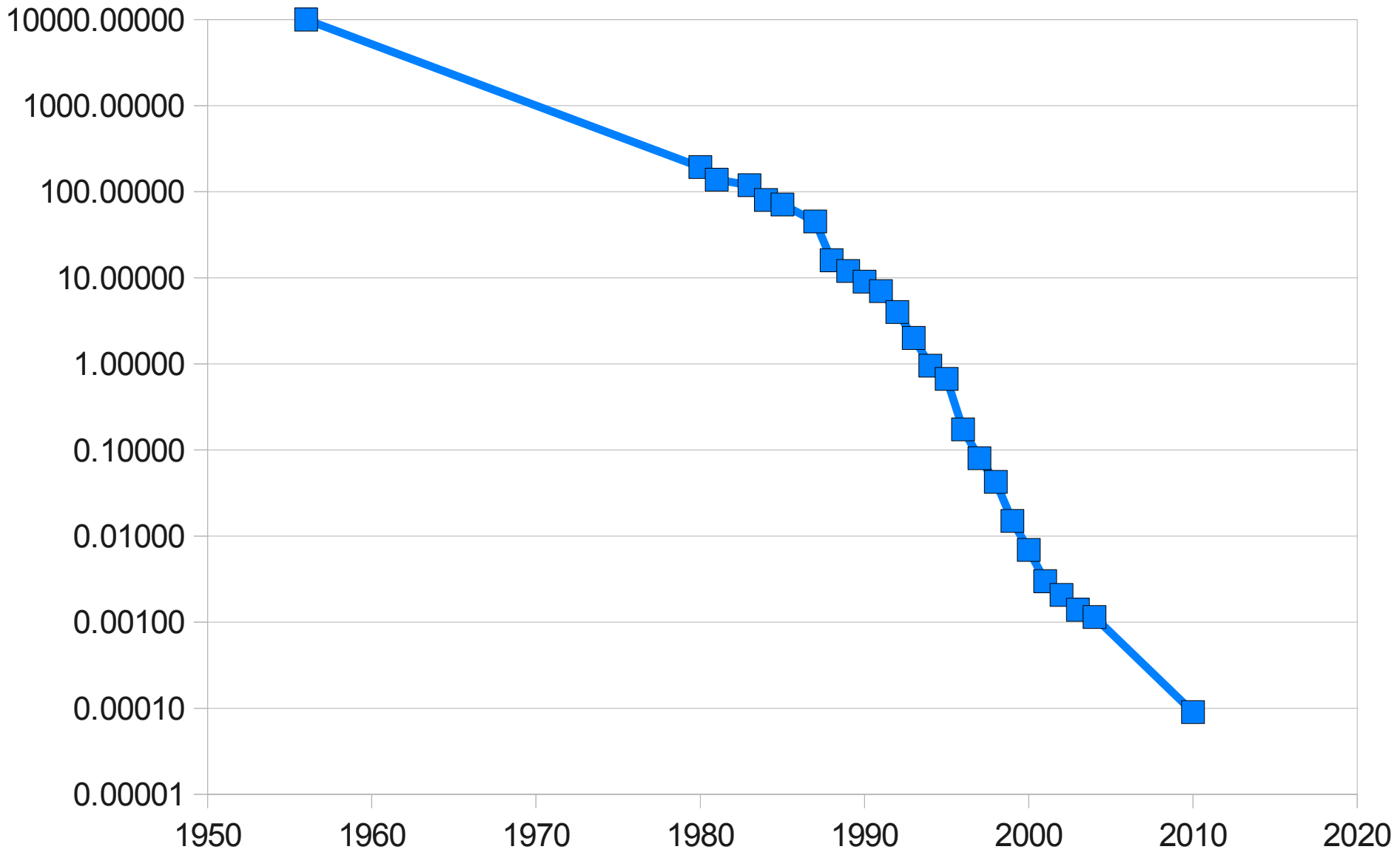


Introdução e Motivação

- Armazenamento do **RAMAC** *(Random Access Method of Accounting and Control)*, IBM, 1956.
- 50 discos de 24 pol. de diâmetro.
- Quase 5 megabytes.
- Custo: Us$ 50.000

Leia mais em http://en.wikipedia.org/wiki/RAMAC e http://www-03.ibm.com/ibm/history/exhibits/650/650_album.html

Introdução e Motivação



http://www.littletechshoppe.com/ns1625/winchest.html

Introdução

1981

www.vintagecomputing.com

2010

Us$ 370/M → Us$ 0.00009/M

Us$ 180.

Aquisição de Dados

- Crescimento explosivo na capacidade de gerar, coletar e armazenar dados:
  - Científicos: imagens, sinais.
  - Sociais: censos, pesquisas.
  - Econômicos e comerciais: transações bancárias e comerciais, compras, ligações telefônicas, acessos à web, transações com código de barras e RFID.
  - Segurança: acessos à sistemas em rede (*logs*), e-mails corporativos, registro de atividades.
- Justificativas para este aumento:
  - Barateamento de componentes e ambientes computacionais.
  - Exigências científicas/sociais.
  - Mudança de paradigmas (em particular na Web)!

Aquisição de Dados (Científicos/Espaciais)

- *Max Planck Institute for Meteorology:* 220 terabytes de dados de pesquisa sobre o clima.
- *LHC: Large Hadron Collider* do CERN: 15 petabytes de dados por ano.
- *SDSS (Sloan Digital Sky Survey):* 40 terabytes de dados (imagens mais catálogo de 200 milhões de objetos mais outros dados).
- *LSST (Large Synoptic Survey Telescope):* meio petabyte de imagens por mês, catálogo de 300 terabytes por ano.
- Microsoft TerraServer: 5 terabytes (1999).
- INPE: 130 terabytes de imagens de sensoriamento remoto.

Aquisição de Dados (Web)

- *CiteSeerX: 1.400.000 artigos científicos, 27.000.000 citações.*
- *Springer: 4.400.000 artigos científicos.*
- *Sourceforge: 230.000 projetos de software aberto.*
- *YouTube*: 45 terabytes de vídeos em 2006.
- *Flickr*: 3.7 bilhões de imagens.
- *Facebook*: 250.000.000 usuários, 45.000.000 grupos de interesse, 1.000.000.000 fotos por mês.
- *Wayback machine*: 2 petabytes, 20 terabytes/mês, 55 bilhões de páginas.

Aquisição de Dados

- Mídia impressa, filmes, mídia magnética e ótica produziram aproximadamente 5 exabytes de novos dados em 2002.
  - 1 exabyte = 1.024 petabytes = 1.048.576 terabytes.
- Consumidor americano típico gera 100G de dados em sua vida:

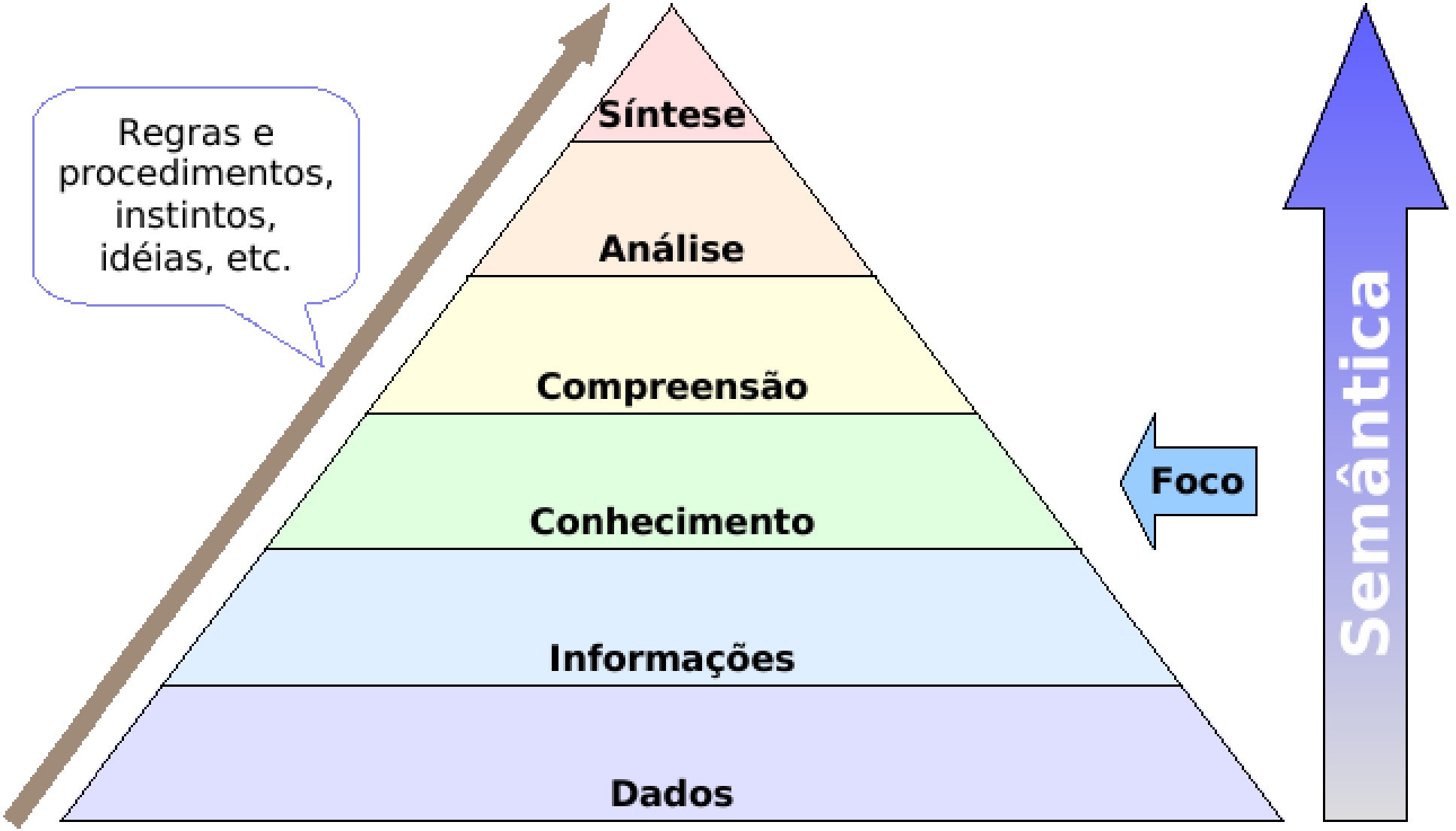
  =~ 26 exabytes para a população presente.
- Quantos registros de ligações telefônicas?
- Quantas transações de cartões por dia?
- Quantos acessos a diversos servidores de informação?

- *O que você tem no seu disco rígido?*

Aquisição de Dados

- Mas o que é feito destes dados? Como “olhar” estes dados?
  - Localizar, filtrar é relativamente simples...
  - Indexar pode ser mais complicado.
- Como identificar..
  - Padrões (“X” acontece se...)
  - Exceções (isto é diferente de... por causa de...)
  - Tendências (ao longo do tempo, “Y” deve acontecer...)
  - Correlações (se “M” acontece, “N” também deve acontecer.)
- O que existe de interessante nestes dados? Como definir “interessante”?
- **Informação**, e não dados, valem dinheiro / tempo / conhecimento!

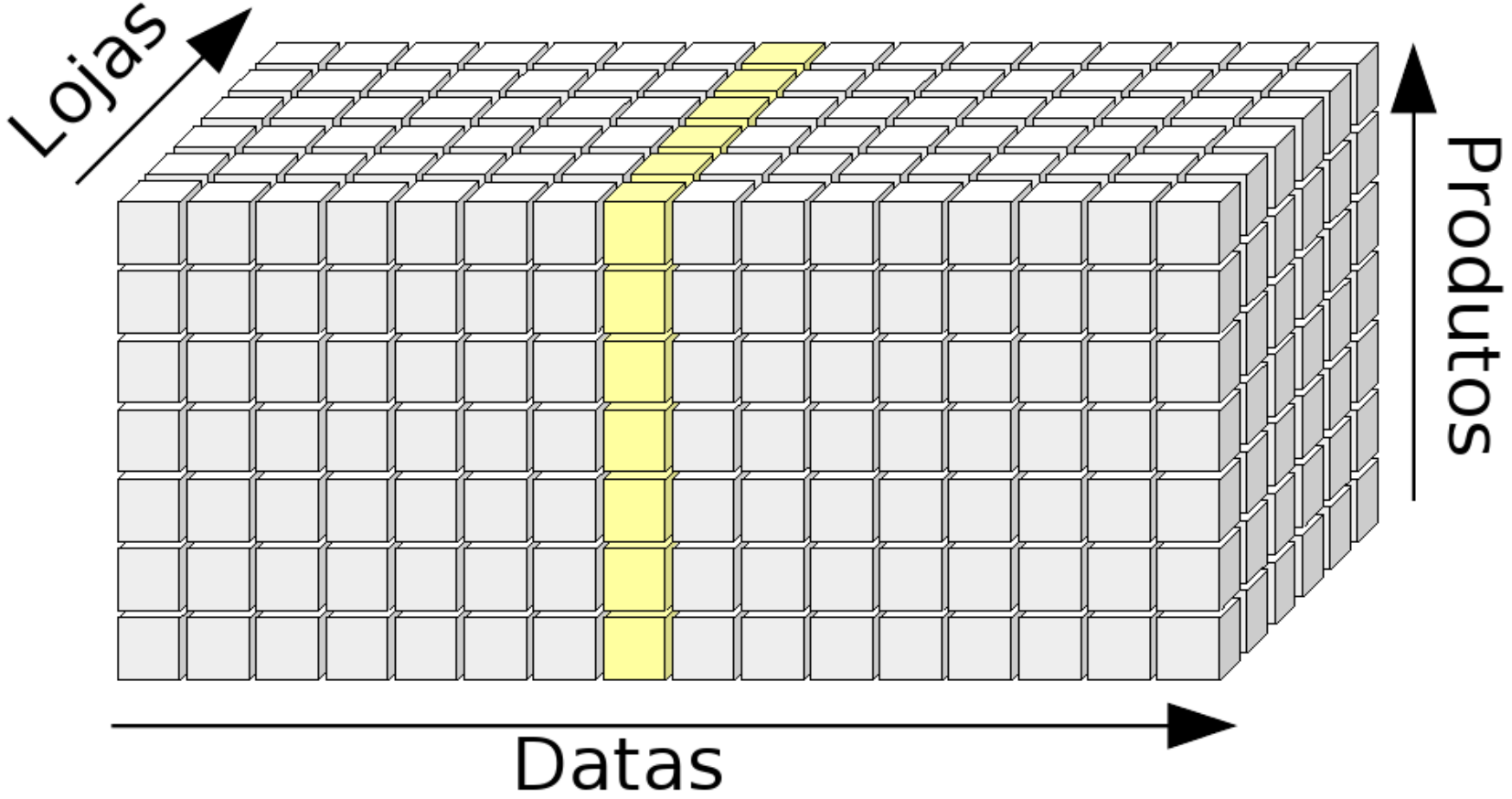
Dados, Informações, Conhecimento



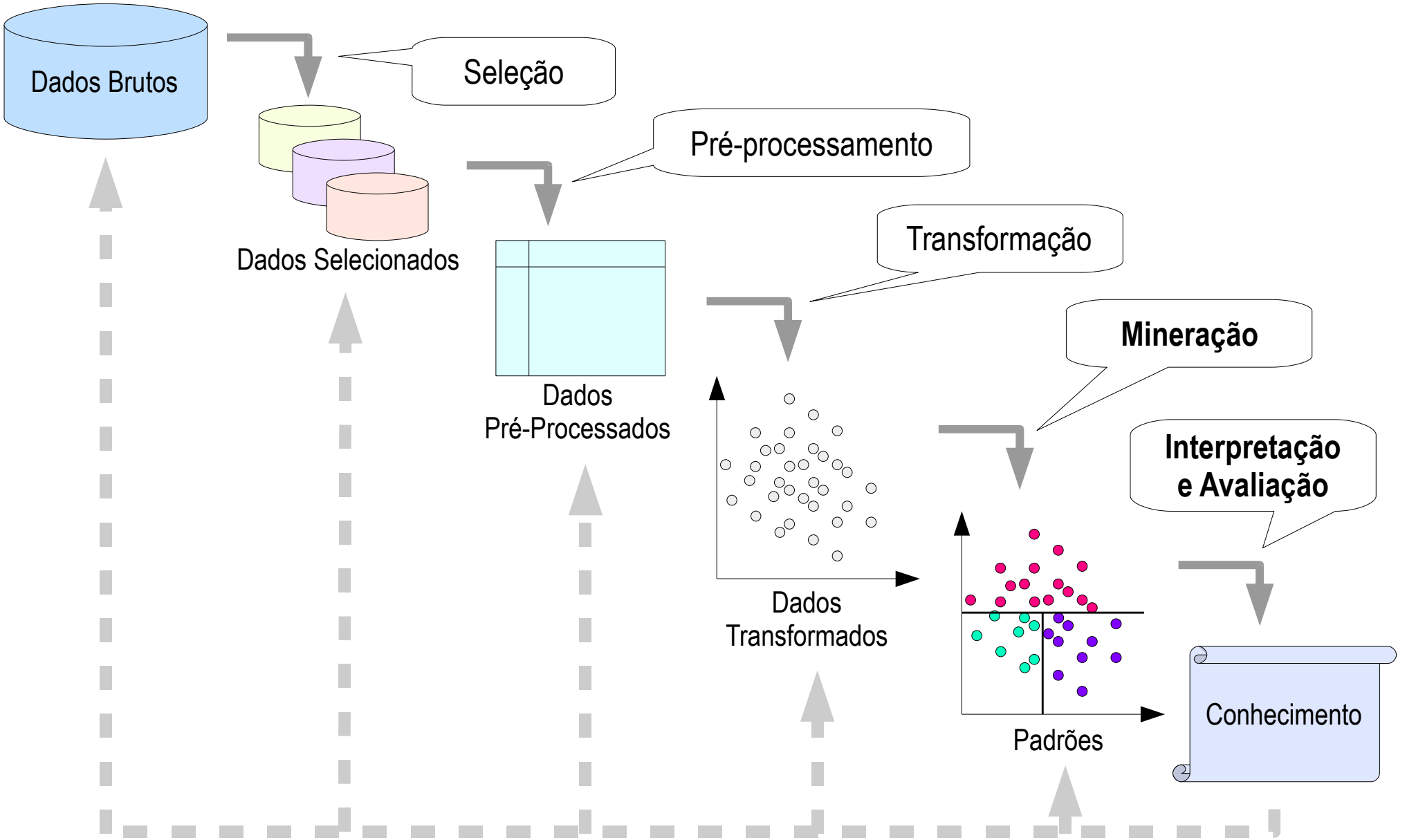
Mineração de Dados

- Parte do processo de descoberta de conhecimentos em bancos de dados (*Knowledge Discovery in Databases, KDD*).
- **KDD**: Processo geral de descoberta de conhecimentos úteis previamente desconhecidos a partir de grandes bancos de dados (adaptado de Fayyad *et al*).

- Não é SQL nem OLAP, embora estas técnicas possam ser parte do processo.



Knowledge Discovery in Databases



Passos do Processo

- De acordo com Fayyad et. al.
  1. Compreensão do domínio da aplicação.
  2. Criação de conjunto de dados para descoberta.
  3. Limpeza e pré-processamento dos dados.
  4. Redução e reprojeção.
  5. Escolha da tarefa de mineração de dados.
  6. Escolha dos algoritmos de mineração e de seus parâmetros.
  7. Mineração de dados.
  8. Interpretação.
  9. Consolidação e avaliação.

Definições

- *Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand, Mannila and Smyth, Principles of Data Mining).*

- *Data mining is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization to address the issue of information extraction from large data bases (Evangelos Simoudis,* citado em *Daniel T. Larose, Discovering Knowledge in Data – An Introduction to Data Mining).*

Definições

- Interseção com outras técnicas e ciências.
- **Não é** a “nova estatística!”
- Usa muitos conceitos e técnicas de estatística, reconhecimento de padrões, aprendizado por máquina, inteligência artificial, bancos de dados, processamento de alto desempenho, visualização, etc.
- Tem caráter exploratório e prático.
- **Não dispensa interação e supervisão humanas!**

Casos (Comerciais)

- **Amazon.com**: melhoria da customização da interface com o usuário (melhoria de vendas por indicação), eliminação de fraudes.
- **1-800-FLOWERS.com**: compreensão e antecipação de comportamento de clientes, descoberta de tendências e explicação de observações (CRM).
- **U.S. Census Bureau**: análise de dados espaciais (com SAS e software da ESRI) de ensino público para determinar políticas para melhoria na educação.
- **Japan Credit Bureau**: melhoria da resposta a campanhas de marketing, retenção de clientes, identificação de novos segmentos de mercado.

SAS Success stories: http://www.sas.com/success/technology.html

Casos (Comerciais)

- **Columbia Interactive/Columbia University**: Análise de visitas a sites, coletando "trilhas" de usuários (como usam o site, que páginas são mais atraentes para usuários, quando usuários deixam o site) para melhorar interatividade e planejar conteúdo.
- **Casino**: cadeia com 115 hipermercados, 400 supermercados, mais de 4000 lojas e 260 lanchonetes. Criou programa de cartões de fidelidade e tem coletado dados dos cartões e hábitos de consumo.
- **TIM (Telecom Italia Mobile)**: redução de *churn*, análise de comportamento do usuário e segmentação do banco de dados de usuários.

SAS Success stories: http://www.sas.com/success/technology.html

Casos (Comerciais)

- **IMS America**: Empresa de pesquisa de mercado farmacêutico, mantém um banco de dados de 1.5 bilhões de prescrições de 600.000 médicos, usadas em 33.000 farmácias. Usa o banco para verificar que médicos mudaram seu padrão de prescrições para informar à companhias farmacêuticas, que podem decidir por campanhas de marketing dirigido aos médicos. **Aparentemente agora estão impedidos legalmente de continuar operando.**
- **Harrah's Entertainment Inc.:** Cassino, dobrou lucros usando informações de cartões de "jogadores freqüentes", identificando que um grupo de jogadores que gastavam entre 100 e 499 dólares (30% dos jogadores) geravam a maior parte do lucro do cassino. Testou diferentes promoções para este grupo, obtendo melhor fidelidade com menor custo e aumentando a resposta a campanhas de marketing.

Miriam Wasserman, *Mining data*. http://www.bos.frb.org/economic/nerr/rr2000/q3/mining.htm

Casos (científicos)

- Muitos artigos nas áreas:
  - Mineração de dados espaciais/espaço-temporais, Análise de objetos móveis e trajetórias.
  - Mineração de imagens e sinais de diversos tipos.
  - Segurança, detecção de intrusão, análise de *logs*, análise de *malware*, *spam* e *worms*.
  - Tráfego e roteamento de redes.
  - Análise de grafos / redes de conexões (ex. redes sociais).
  - Análise de documentos (XML, HTML).
  - Bioinformática.

Casos (sem sucesso)

- Evidentemente raros e não anunciados...
  - ***Total Information Awareness***: forte rejeição pela ACLU, outras entidades.
  - **Gazelle.com**: caso-teste, investimento não seria recuperado.
  - Bebidas dietéticas levam a obesidade.
  -
- Muitos dos esforços de mineração de dados resultam em informações pouco úteis!
  - Mas podem aumentar o conhecimento sobre o processo como um todo!

Falácias de *Data Mining*

- ***Data Mining* é automático:** é um processo, é iterativo, requer supervisão.
- **Investimentos são recuperados rapidamente:** depende de muitos fatores!
- ***Software* são *intuitivos e simples:*** é mais importante conhecer os conceitos dos algoritmos e o negócio em si!
- ***Data Mining* pode identificar problemas no negócio:** DM pode encontrar padrões e fenômenos, identificar causa deve ser feito por especialistas.

Adaptado de Daniel T. Larose, ***Discovering Knowledge in Data – An Introduction to Data Mining***

Analogia

Ensinando *Data Mining*

- Falamos sobre terabytes e petabytes, mas não podemos mostrar exemplos práticos nesta escala.
- Falamos sobre dezenas ou centenas de atributos de diversos tipos, mas não é simples demonstrar algoritmos usando-os.
- Ficamos limitados a *toy problems*, geralmente em duas dimensões numéricas, focando mais em características do algoritmo do que em performance e escalabilidade.

Conceitos Básicos

Conceitos Básicos

- Um exemplo (quase) prático.
- Categorias de algoritmos de mineração de dados.
- Representação de dados para mineração de dados.
  - Tipos de atributos.
- Espaço de Atributos.
- Pré-processamento.

Exemplo (quase) prático

Instâncias

Atributos

k	A_1	A_2	A_3	A_4	A_5	classe
1	010	0.60	0.70	1.7	I	alto
2	060	0.70	0.60	1.3	P	alto
3	100	0.40	0.30	1.8	P	médio
4	120	0.20	0.10	1.3	P	médio
5	130	0.45	0.32	1.9	I	baixo
6	090	?	0.18	2.2	I	?
7	110	0.45	0.22	1.4	P	?

Exemplo (quase) prático

k	A_1	A_2	A_3	A_4	A_5	classe
1	010	0.60	0.70	1.7	I	alto
2	060	0.70	0.60	1.3	P	alto
3	100	0.40	0.30	1.8	P	médio
4	120	0.20	0.10	1.3	P	médio
5	130	0.45	0.32	1.9	I	baixo
6	090	?	0.18	2.2	I	?
7	110	0.45	0.22	1.4	P	?

- Existe algum padrão? Existe algo fora de um padrão?
- Quais atributos influenciam nas classes?
  - Podemos escolher a classe em função dos valores dos atributos?
- Podemos prever o valor de um atributo em função de outros?

Técnicas (geral)

- **Classificação:** aprendizado de uma função que pode ser usada para mapear dados em uma de várias classes discretas definidas previamente.
  - A classe é **alto** se $A_1 < 70$ e $A_2 > 0.5$.
- **Regressão ou Predição:** aprendizado de uma função que pode ser usada para mapear os valores associados aos dados em um ou mais valores reais.
  - A_3 pode ser calculado em função de A_2?

Técnicas (geral)

- **Agrupamento (ou *clustering*):** identificação de grupos de dados onde os dados tem características semelhantes aos do mesmo grupo e onde os grupos tenham características diferentes entre si.
- **Sumarização**: descrição do que caracteriza um conjunto de dados (ex. conjunto de regras que descreve o comportamento e relação entre os valores dos dados).

Técnicas (geral)

- **Detecção de desvios ou *outliers*:** identificação de dados que deveriam seguir um padrão esperado mas não o fazem.
- **Identificação de associações:** identificação de grupos de dados que apresentam co-ocorrência entre si (ex. cesta de compras).

- Técnicas podem ser usadas em mais de uma fase do processo de KDD.

Conceitos Básicos: Representação

k	A_1	A_2	A_3	A_4	A_5	classe
1	010	0.60	0.70	1.7	I	alto
2	060	0.70	0.60	1.3	P	alto
3	100	0.40	0.30	1.8	P	médio
4	120	0.20	0.10	1.3	P	médio
5	130	0.45	0.32	1.9	I	baixo
6	090	?	0.18	2.2	I	?
7	110	0.45	0.22	1.4	P	?

- Para facilitar...
  - Dados em uma única tabela.
  - Cada linha na tabela é uma **instância** ou **amostra** (registros).
  - Cada coluna na tabela é um **atributo** (campos).
  - Cada instância da base de dados tem os mesmos campos e que cada campo tem o mesmo tipo de valor.
  - Eventualmente um atributo para uma instância pode ser desconhecido ou estar faltando.

Conceitos Básicos: Atributos

- Tipos de atributos
  - Atributos **nominais** são rótulos, nomes, basicamente servem para identificar uma amostra e diferenciá-la de outra.
  - Atributos **categóricos** são semelhantes aos nominais mas são escolhidos de um conjunto definido.
  - Atributos **numéricos** expressam algo medido (com instrumentos, por exemplo).
  - Atributos **ordinais** são valores discretos mas que apresentam uma ordem imposta ou implícita.
- Podemos transformar alguns tipos em outros.
- Entender a diferença e limitações é muito importante!

Conceitos Básicos: Pré-Processamento

- Pré-Processamento
  - Atributos com representação inadequada para tarefa e algoritmo.
  - Atributos cujos valores não tenham informações adequadas.
  - Excesso de atributos (podem ser redundantes ou desnecessários).
  - Atributos insuficientes.
  - Excesso de instâncias (afetam tempo de processamento).
  - Instâncias insuficientes.
  - Instâncias incompletas (sem valores para alguns atributos).
- Assim como a mineração de dados em si, requer conhecimento sobre os dados e algoritmo que será usado!

Pré-Processamento: Representação Inadequada

- Problemas:
  - Redes Neurais *Back-propagation* só operam com valores numéricos.
  - Alguns algoritmos de busca de associações só operam com valores simbólicos/discretos.
- Soluções:
  - Conversão de tipos de atributos (quando aplicável!)
  - Remoção dos atributos inadequados.
  - Separação em subtarefas usando os valores discretos dos atributos.

Pré-Processamento: Valores Inadequados

- Problemas:
  - Atributos com baixíssima variabilidade nos valores.
  - Atributos redundantes ou altamente correlacionados com outros.
- Soluções:
  - Remoção dos atributos inadequados.
  - Unificação de atributos ou derivação de novos atributos.

Pré-Processamento: Excesso de Atributos

- Problemas:
  - Muitos atributos ⟶ complexidade de processamento.
  - Correlações irrelevantes podem complicar o processo de mineração (a não ser que seja necessário descobri-las!)
- Soluções:
  - Remoção dos atributos irrelevantes (possivelmente depois de alguma análise).
  - Mudança de representação ou projeção (usando, por exemplo, *PCA* ou Mapas de Kohonen).

Pré-Processamento: Atributos Insuficientes

- Problemas:
  - Poucos atributos podem não possibilitar mineração adequada (para identificar classes, por exemplo).
- Soluções:
  - Enriquecimento com dados complementares (se puderem ser obtidos!)
  - Enriquecimento com combinações não lineares.
  - *Data Farming*.

Criação de grandes volumes de dados através de simulação com alta performance usando modelos generativos de dados.

Pré-Processamento: Redução de Instâncias

- Problemas:
  - Muitas instâncias podem tornar o processamento inviável: alguns algoritmos requerem várias iterações com os dados.
  - Problema relacionado: desbalanceamento de instâncias para classificação.
- Soluções:
  - Redução por amostragem.
  - Redução por prototipagem.
  - Particionamento do conjunto de dados.

Pré-Processamento: Instâncias Insuficientes

- Problemas:
  - Poucas instâncias podem comprometer o resultado (que será pouco genérico ou confiável).
  - Casos raros podem não ser representados.
- Soluções:
  - Coleta de mais instâncias.
  - *Data Farming*.

Pré-Processamento: Dados Incompletos

- Problemas:
  - Dados coletados podem ter valores de atributos faltando.
  - Por que estão faltando? Rever modelagem do processo e coleta!
- Soluções:
  - Eliminação de dados/atributos com muitos valores faltando.
  - Completar através de proximidade/similaridade com dados completos.
  - Separar em conjuntos para processamento independente ou associado.

Pré-Processamento: Para quê?

- Restrições dos algoritmos (para aplicabilidade, para garantir completeza e para reduzir complexidade).
  - É possível/viável?
- Devemos também considerar...
  - Atributos e dados podem/devem ser representados de outra forma?
  - Algumas conversões de tipos podem ser destrutivas: cuidado com discretização!

k	A_1	A_2	A_3	A_4	A_5	classe
1	010	0.60	0.70	1.7	I	alto
2	060	0.70	0.60	1.3	P	alto
3	100	0.40	0.30	1.8	P	médio
4	120	0.20	0.10	1.3	P	médio
5	130	0.45	0.32	1.9	I	baixo
6	090	?	0.18	2.2	I	?
7	110	0.45	0.22	1.4	P	?

Conceitos Básicos: Espaço de Atributos

- Instâncias são vetores de dados em um espaço N-dimensional.
  - Que "aparência" tem a distribuição das instâncias no espaço de atributos?
  - Existe correlação entre atributos?
  - Existe possibilidade de classificação simples?
  - Existem desvios ou *outliers* comprometedores?
  - As classes implícitas nos dados são separáveis?
- Conceito de *proximidade no espaço N-dimensional* (= **semelhança** de atributos) essencial!

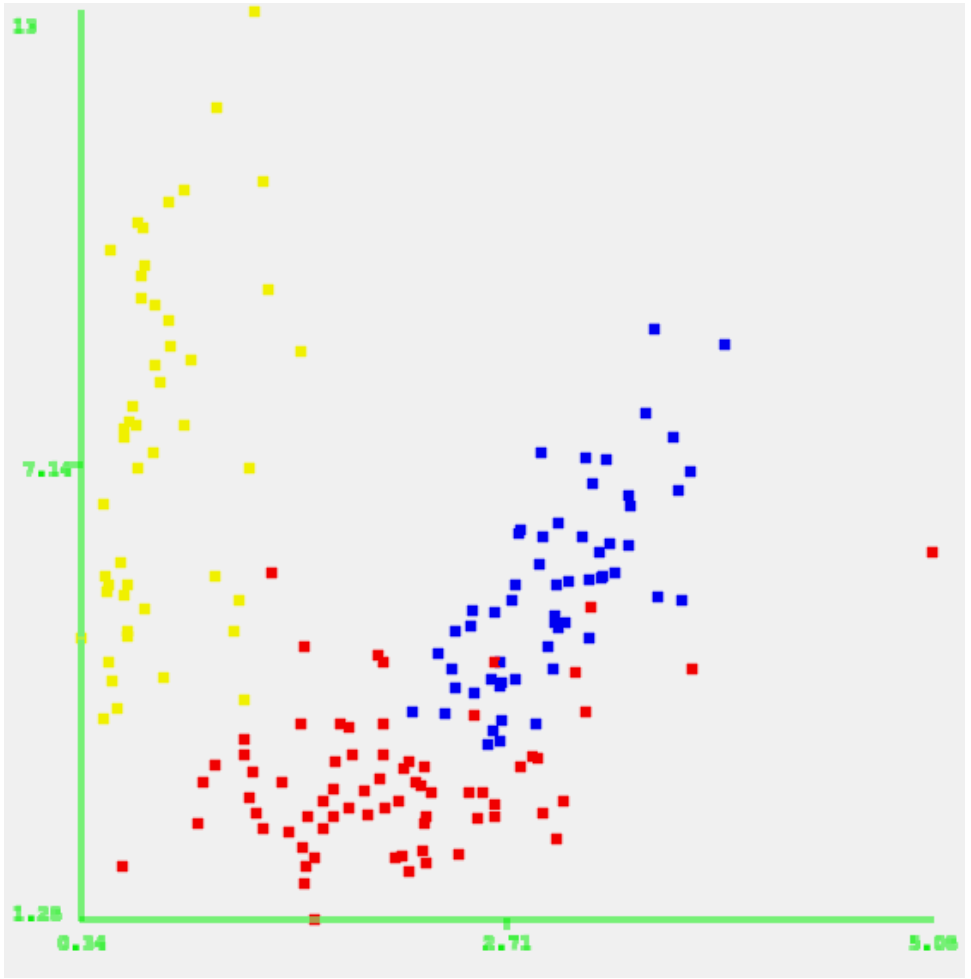
Conceitos Básicos: Espaço de Atributos

- Origem do vinho a partir de conteúdo físico-químico (13 atributos)
  http://archive.ics.uci.edu/ml/datasets/Wine (nomes de atributos originais)

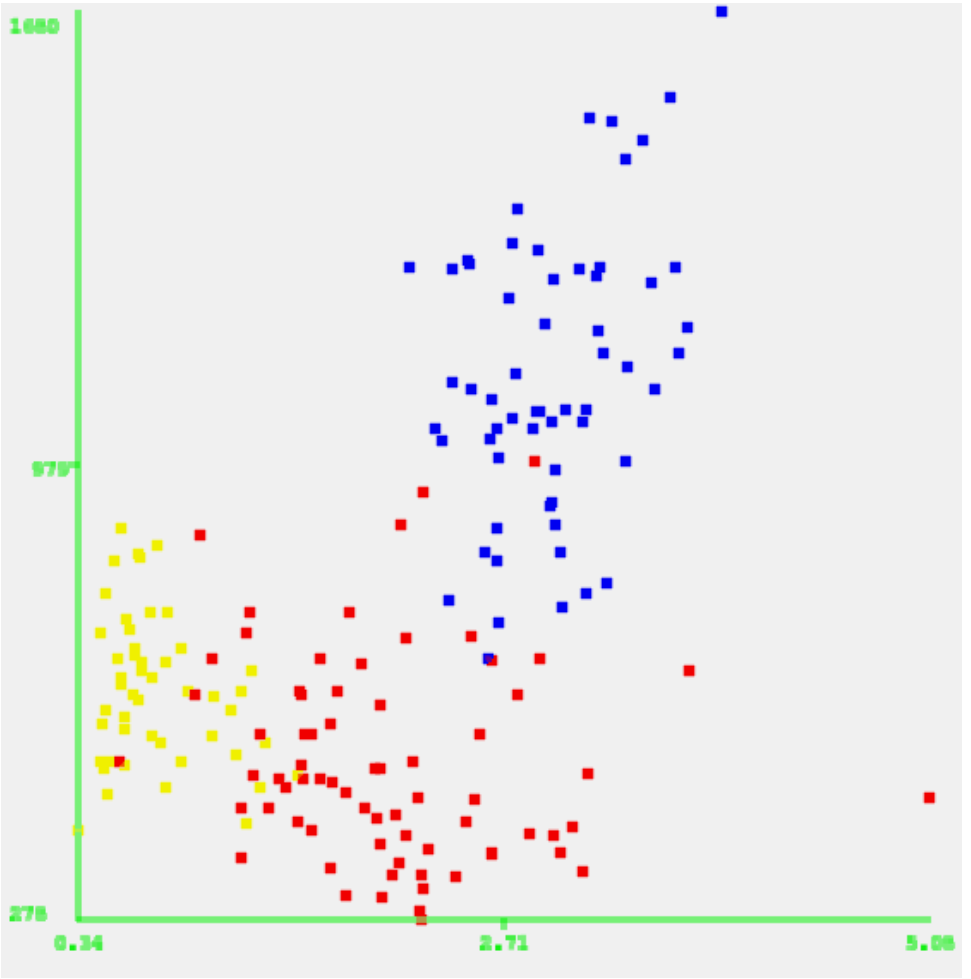
No.	Alcohol Numeric	MalicAcid Numeric	Ash Numeric	AlcalinityOfAsh Numeric	Magnesium Numeric	TotalPhenols Numeric	Flavanoids Numeric	NonflavanoidPhenols Numeric	Proanthocyanins Numeric	ColorIntensity Numeric	Hue Numeric	OD280_OD315OfDilutedWines Numeric	Proline Numeric	**ORIGIN** Nominal
1	14.23	1.71	2.43	15.6	127.0	2.8	3.06	0.28	2.29	5.64	1.04	3.92	106...	1
2	13.2	1.78	2.14	11.2	100.0	2.65	2.76	0.26	1.28	4.38	1.05	3.4	105...	1
3	13.16	2.36	2.67	18.6	101.0	2.8	3.24	0.3	2.81	5.68	1.03	3.17	118...	1
4	14.37	1.95	2.5	16.8	113.0	3.85	3.49	0.24	2.18	7.8	0.86	3.45	148...	1
5	13.24	2.59	2.87	21.0	118.0	2.8	2.69	0.39	1.82	4.32	1.04	2.93	735.0	1
6	14.2	1.76	2.45	15.2	112.0	3.27	3.39	0.34	1.97	6.75	1.05	2.85	145...	1
7	14.39	1.87	2.45	14.6	96.0	2.5	2.52	0.3	1.98	5.25	1.02	3.58	129...	1
8	14.06	2.15	2.61	17.6	121.0	2.6	2.51	0.31	1.25	5.05	1.06	3.58	129...	1
9	14.83	1.64	2.17	14.0	97.0	2.8	2.98	0.29	1.98	5.2	1.08	2.85	104...	1
10	13.86	1.35	2.27	16.0	98.0	2.98	3.15	0.22	1.85	7.22	1.01	3.55	104...	1
11	14.1	2.16	2.3	18.0	105.0	2.95	3.32	0.22	2.38	5.75	1.25	3.17	151...	1
12	14.12	1.48	2.32	16.8	95.0	2.2	2.43	0.26	1.57	5.0	1.17	2.82	128...	1
13	13.75	1.73	2.41	16.0	89.0	2.6	2.76	0.29	1.81	5.6	1.15	2.9	132...	1
14	14.75	1.73	2.39	11.4	91.0	3.1	3.69	0.43	2.81	5.4	1.25	2.73	115...	1
15	14.38	1.87	2.38	12.0	102.0	3.3	3.64	0.29	2.96	7.5	1.2	3.0	154...	1
16	13.63	1.81	2.7	17.2	112.0	2.85	2.91	0.3	1.46	7.3	1.28	2.88	131...	1
17	14.3	1.92	2.72	20.0	120.0	2.8	3.14	0.33	1.97	6.2	1.07	2.65	128...	1
18	13.83	1.57	2.62	20.0	115.0	2.95	3.4	0.4	1.72	6.6	1.13	2.57	113...	1
19	14.19	1.59	2.48	16.5	108.0	3.3	3.93	0.32	1.86	8.7	1.23	2.82	168...	1
20	13.64	3.1	2.56	15.2	116.0	2.7	3.03	0.17	1.66	5.1	0.96	3.36	845.0	1
21	14.06	1.63	2.28	16.0	126.0	3.0	3.17	0.24	2.1	5.65	1.09	3.71	780.0	1
22	12.93	3.8	2.65	18.6	102.0	2.41	2.41	0.25	1.98	4.5	1.03	3.52	770.0	1
23	13.71	1.86	2.36	16.6	101.0	2.61	2.88	0.27	1.69	3.8	1.11	4.0	103...	1
24	12.85	1.6	2.52	17.8	95.0	2.48	2.37	0.26	1.46	3.93	1.09	3.63	101...	1
25	13.5	1.81	2.61	20.0	96.0	2.53	2.61	0.28	1.66	3.52	1.12	3.82	845.0	1
26	13.05	2.05	3.22	25.0	124.0	2.63	2.68	0.47	1.92	3.58	1.13	3.2	830.0	1
27	13.39	1.77	2.62	16.1	93.0	2.85	2.94	0.34	1.45	4.8	0.92	3.22	119...	1
28	13.3	1.72	2.14	17.0	94.0	2.4	2.19	0.27	1.35	3.95	1.02	2.77	128...	1
29	13.87	1.9	2.8	19.4	107.0	2.95	2.97	0.37	1.76	4.5	1.25	3.4	915.0	1
30	14.02	1.68	2.21	16.0	96.0	2.65	2.33	0.26	1.98	4.7	1.04	3.59	103...	1
31	13.73	1.5	2.7	22.5	101.0	3.0	3.25	0.29	2.38	5.7	1.19	2.71	128...	1
32	13.58	1.66	2.36	19.1	106.0	2.86	3.19	0.22	1.95	6.9	1.09	2.88	151...	1
33	13.68	1.83	2.36	17.2	104.0	2.42	2.69	0.42	1.97	3.84	1.23	2.87	990.0	1
34	13.76	1.53	2.7	19.5	132.0	2.95	2.74	0.5	1.35	5.4	1.25	3.0	123...	1
35	13.51	1.8	2.65	19.0	110.0	2.35	2.53	0.29	1.54	4.2	1.1	2.87	109...	1
36	13.48	1.81	2.41	20.5	100.0	2.7	2.98	0.26	1.86	5.1	1.04	3.47	920.0	1
37	13.28	1.64	2.84	15.5	110.0	2.6	2.68	0.34	1.36	4.6	1.09	2.78	880.0	1
38	13.05	1.65	2.55	18.0	98.0	2.45	2.43	0.29	1.44	4.25	1.12	2.51	110...	1
39	13.07	1.5	2.1	15.5	98.0	2.4	2.64	0.28	1.37	3.7	1.18	2.69	102...	1

Conceitos Básicos: Espaço de Atributos

- "Olhando" os dados!

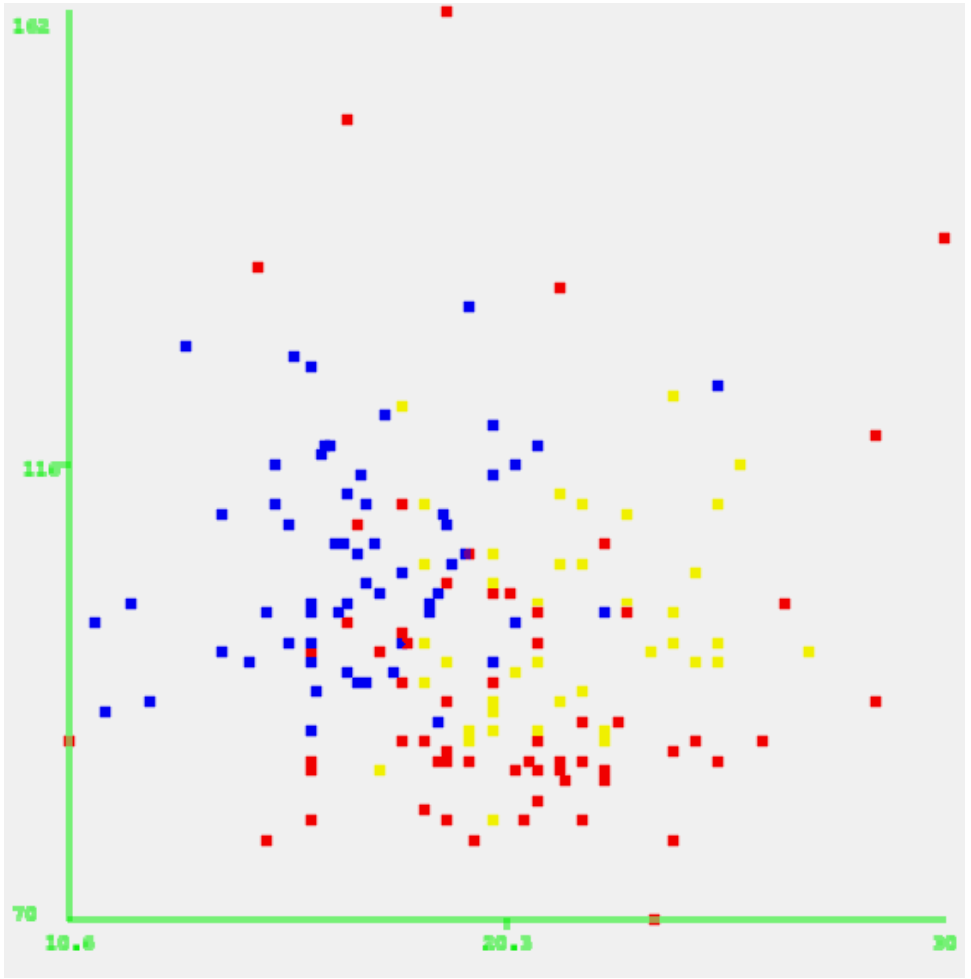


X: Flavonoids, Y: Color Intensity

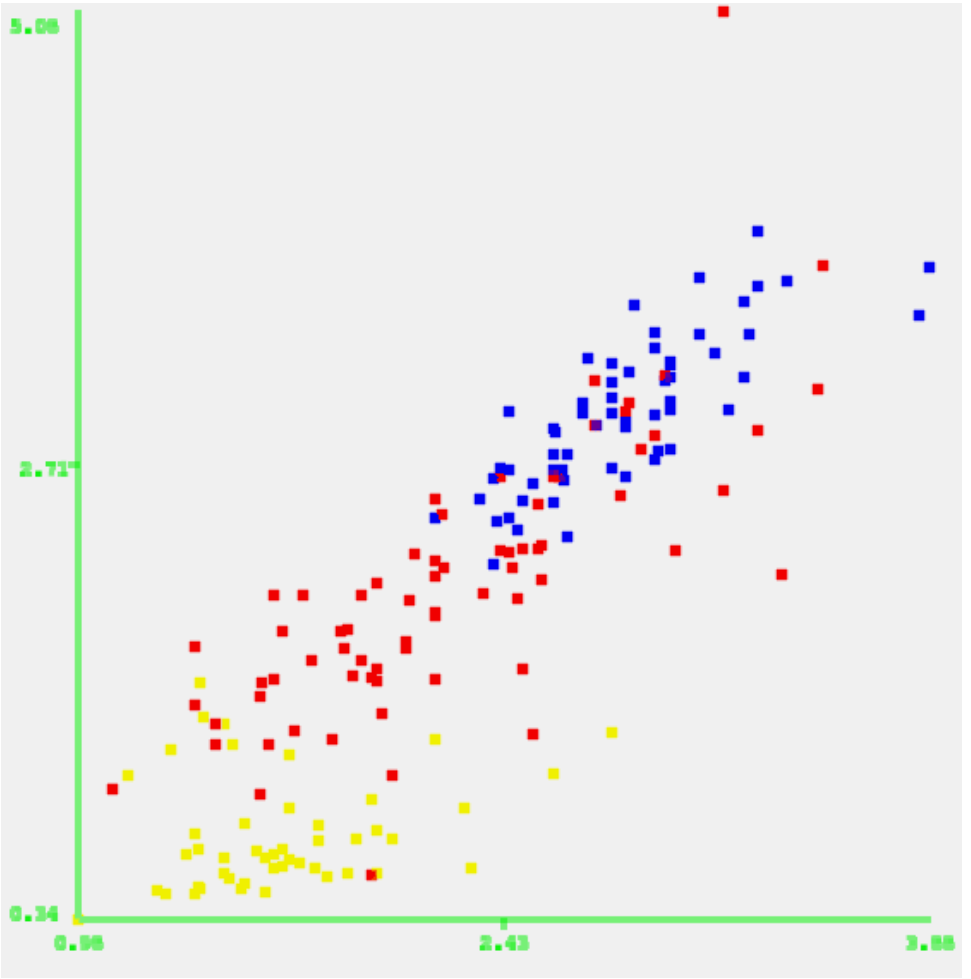


X: Flavonoids, Y: Proline

Conceitos Básicos: Espaço de Atributos



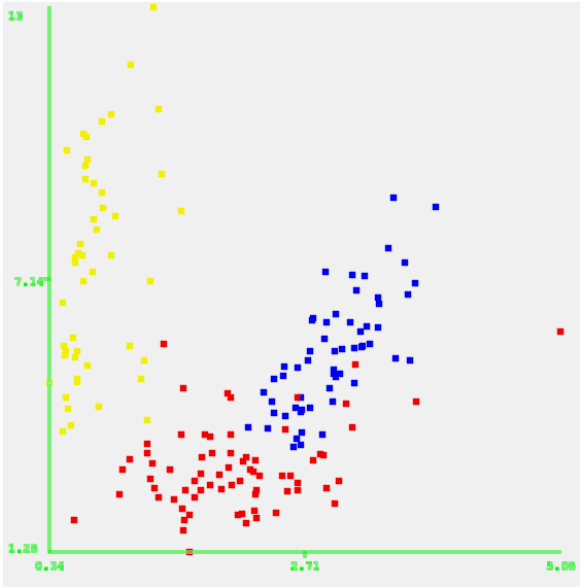
X: Alkalinity of Ash, Y: Magnesium



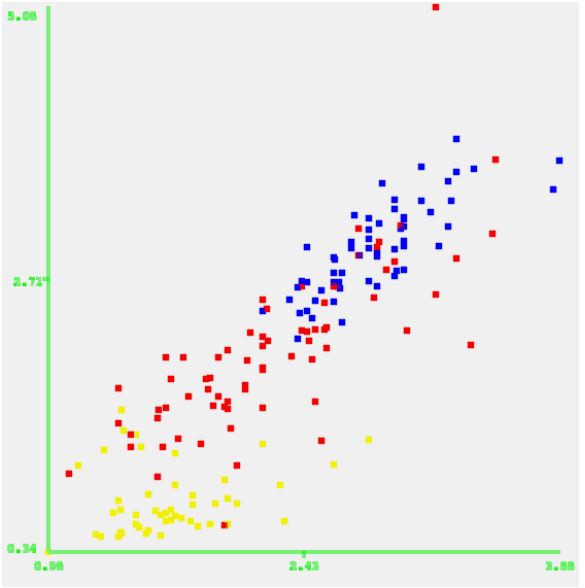
X: Total Phenols, Y: Flavonoids

Conceitos Básicos: Espaço de Atributos

- Visualização pode mostrar várias informações sobre os dados!
  - Quais atributos permitem separação em classes?
  - Quais atributos são correlacionados?
  - Como é a distribuição das classes (se houver)?
  - Existem estruturas interessantes?

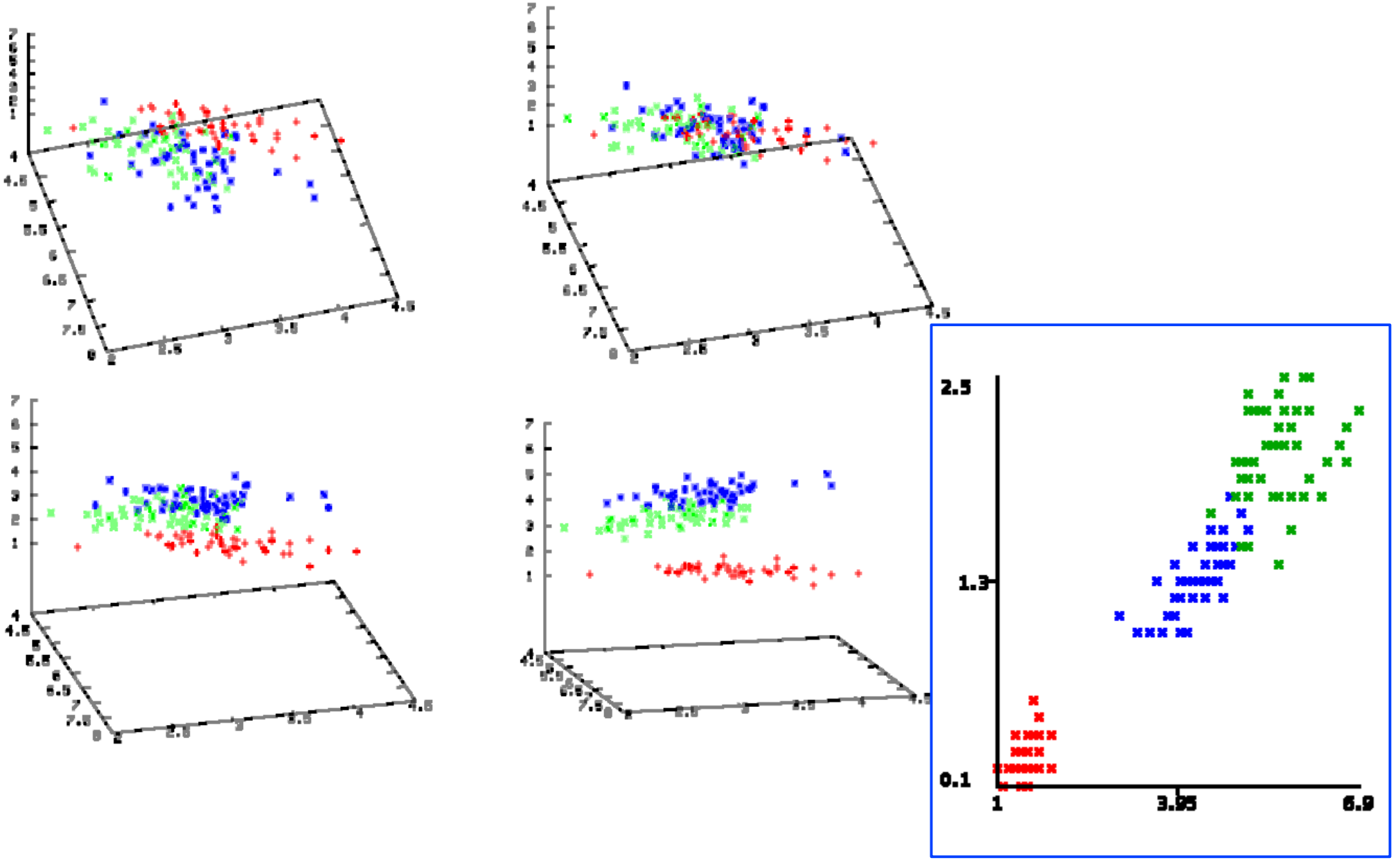


X: Flavonoids, Y: Color Intensity



X: Total Phenols, Y: Flavonoids

Conceitos Básicos: Espaço de Atributos



Próximos passos...

- ***Dia 1:*** Apresentação dos conceitos de mineração de dados, motivação e alguns exemplos.
- ***Dia 2:*** Algoritmos de classificação supervisionada e aplicações.
- ***Dia 3:*** Algoritmos de classificação não-supervisionada e aplicações. Algoritmos de mineração de associações.
- ***Dia 4:*** Visualização e mineração de dados. Outros algoritmos e idéias. Onde aprender mais.

Mais informações em...

- http://www.lac.inpe.br/~rafael.santos
  - http://www.lac.inpe.br/~rafael.santos/dmapresentacoes.jsp
  - http://www.lac.inpe.br/~rafael.santos/cap359-2010.jsp
- http://www.lac.inpe.br/ELAC/index.jsp