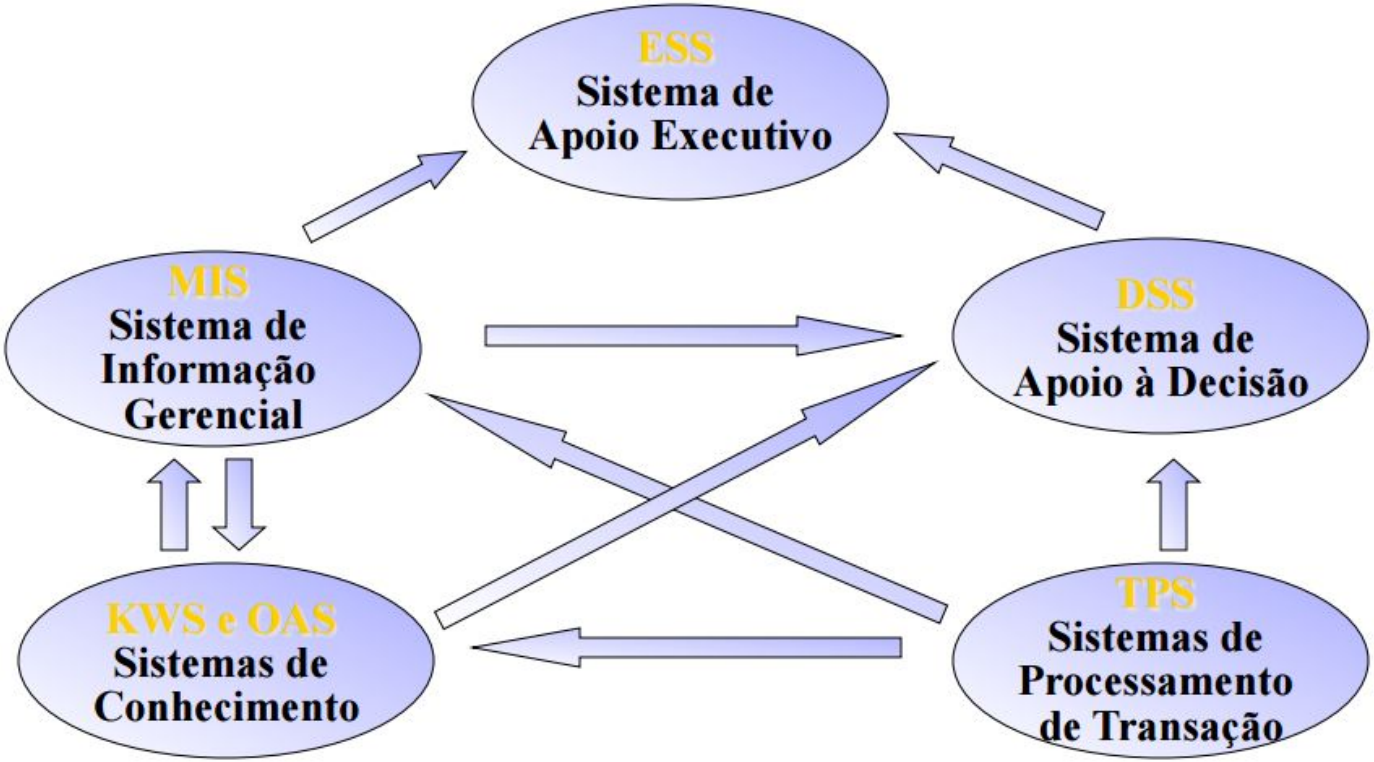
Data Mining e Data Warehouse

Ederson Tyiuji Noya
Guilherme de Freitas Perinazzo
Guilherme Masao Oyakawa
Rafael Silva de Milha

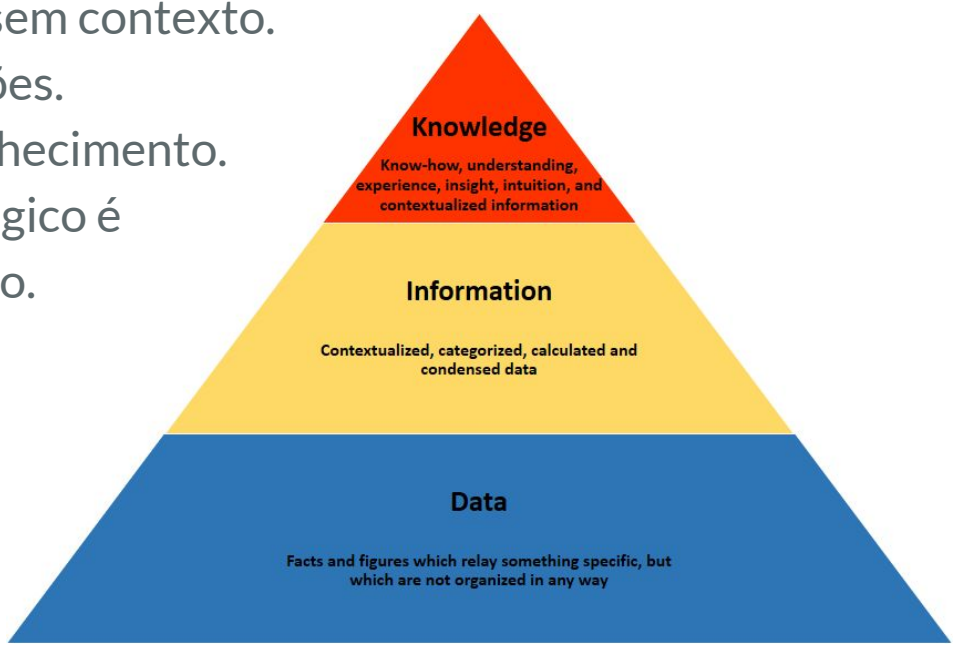
Índice

Tipos de Sistemas de Informação



Revisão: Dados e Sistemas de Informação

- Dados: fatos em estado primário, sem contexto.
- Dados são refinados em informações.
- Informações são refinadas em conhecimento.
- Conhecimento científico e tecnológico é desejável por qualquer organização.

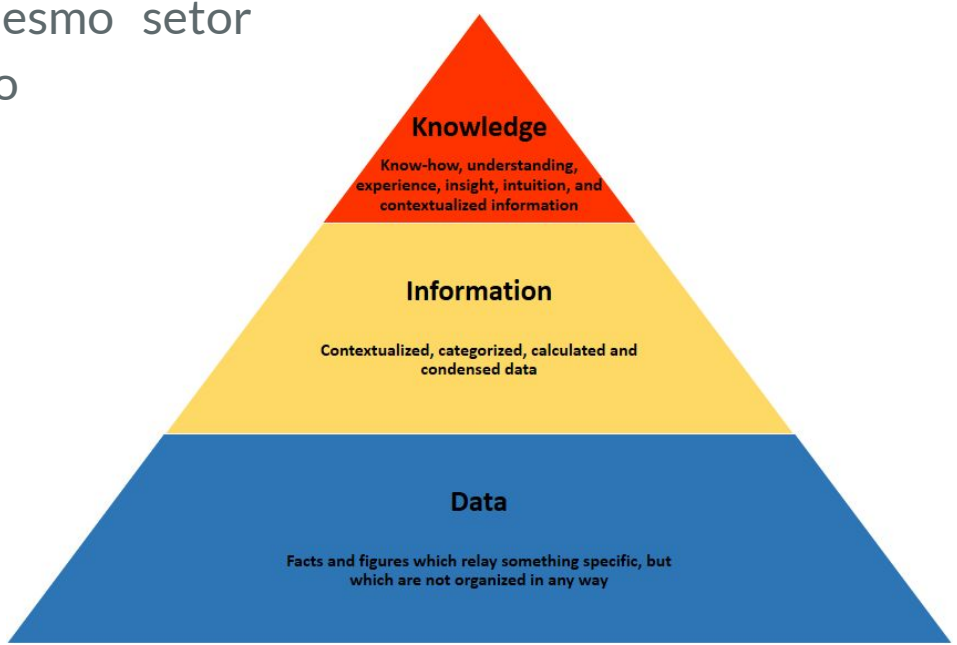


Disponível em http://www.knowledge-management-tools.net/images/Knowledge_pyramid.png

Revisão: Dados e Sistemas de Informação

- Organizações que atuam em um mesmo setor competem entre si pelo domínio do conhecimento.

- Dados são a base de qualquer SI.



Disponível em http://www.knowledge-management-tools.net/images/Knowledge_pyramid.png

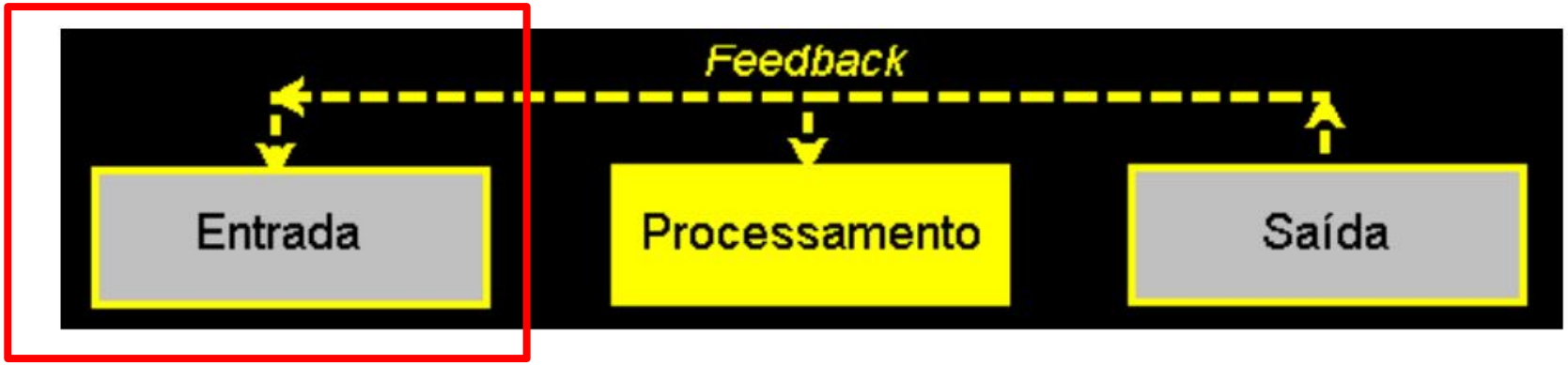
Pipeline de um Sistema de Informação

- Entradas: Coleta de dados (internos à organização ou externos).
- Processamento: Dados → Informação.
- Saída: Distribuição da informação gerada.
- Feedback: Correção do estágio de entrada a partir das saídas.

Pipeline de um Sistema de Informação

- Era da Informação / Era Digital: economia governada por conhecimento.
  - Inovações tecnológicas levam ao crescimento exponencial da geração de dados.

- Necessidade constante e crescente de coleta e processamento de dados.



Compartilhamento de dados entre organizações

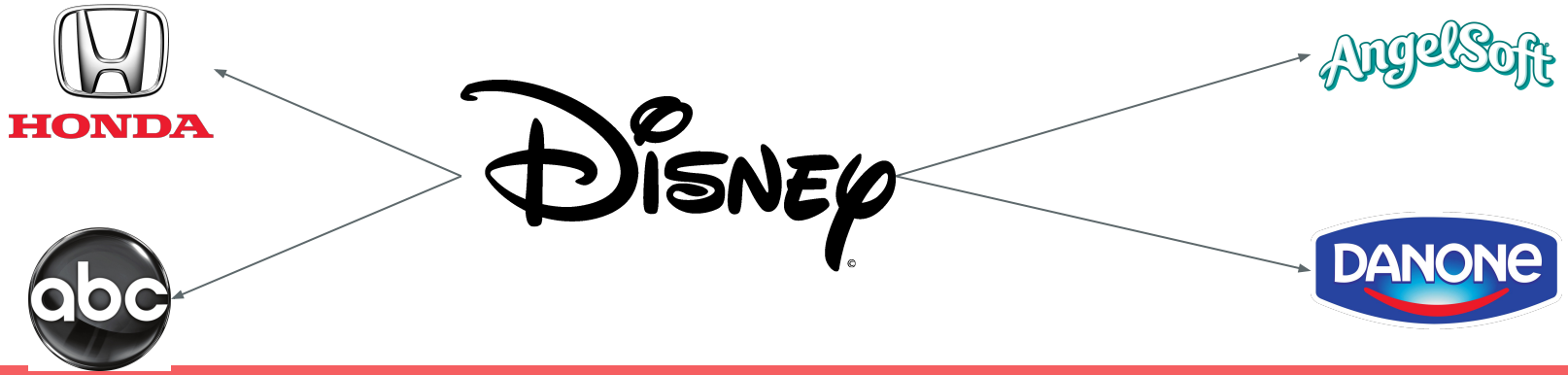
- Dados são tão importantes que atualmente existem empresas dedicadas exclusivamente à captura e venda de dados de consumidores.

- Exemplo: Acxiom Corporation
  - Possui dados de 500 milhões de usuários da Internet
  - Aproximadamente 1500 registros (características) por usuário
  - 2013: investigada pelo FTC pela falta de transparência do processo de coleta e uso de dados

Compartilhamento de dados entre organizações

O que a Disney faz com os dados de seus clientes?

- Pesquisa em 2007 revela que a empresa compartilha informações de usuários cadastrados com organizações associadas
  - Nome, endereço, idade, quantidade de filhos e idade de cada um, ocupação, telefones, produtos comprados...
  - Alguns dos associados: ABC, Honda, Angelsoft, Danone...



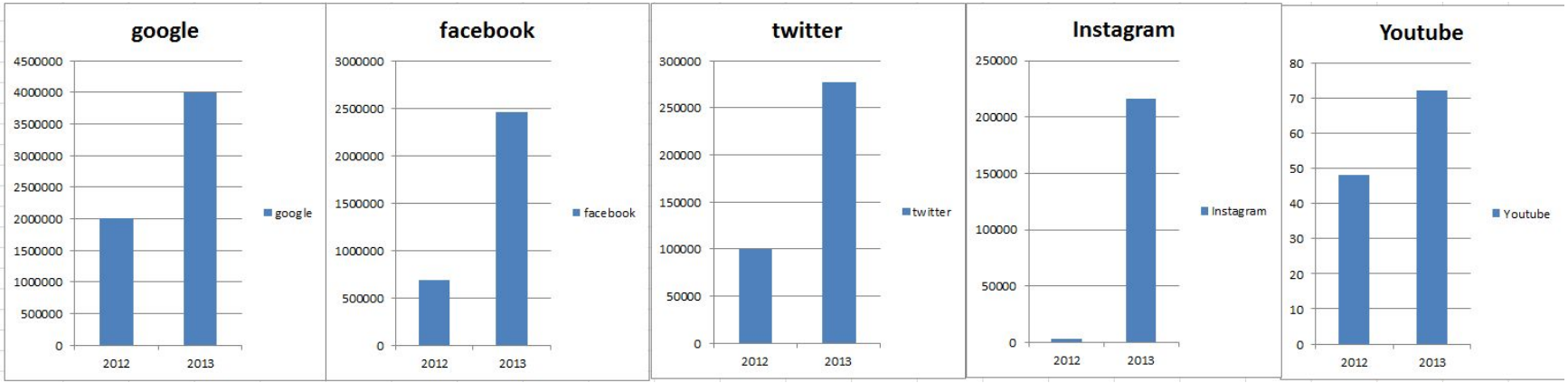
Quantidade de dados gerados na Internet

- Em 2012: 2.1 bilhões de pessoas na Internet.
- A cada minuto: (Fonte: DOMO.com)
  - Mais de 2.000.000 buscas no Google
  - 684.478 posts novos no Facebook
  - Mais de 100.000 tweets
  - Amazon lucra US$66.240 em vendas on-line (valor estimado)
  - 48 horas de vídeo no YouTube
  - 3.600 novas fotos no Instagram
- 2,5 exabytes (10^18 bytes) de dados gerados por dia.

Quantidade de dados gerados na Internet

- Em 2013: 2.4 bilhões de pessoas na Internet.
- A cada minuto: (Fonte: DOMO.com)
  - Mais de 4.000.000 buscas no Google
  - 2.460.000 posts novos no Facebook
  - Mais de 277.000 tweets
  - Amazon lucra US$83.000 em vendas on-line (valor estimado)
  - 72 horas de vídeo no YouTube
  - 216.000 novas fotos no Instagram
- 5 exabytes (10^18 bytes) de dados gerados por dia.
- Tecnologias tradicionais não conseguem trabalhar com esta quantidade de dados.
- Solução: Big Data

Quantidade de dados gerados na Internet



O que é Big Data?

*"Big Data são ativos de informação de alto **volume**, **velocidade**, e/ou **variedade** que requerem novas formas de processamento para permitir melhores tomadas de decisões, descoberta de tendências e otimização de processos." (Tradução Livre)*

-LANEY, Douglas (Gartner), 2012

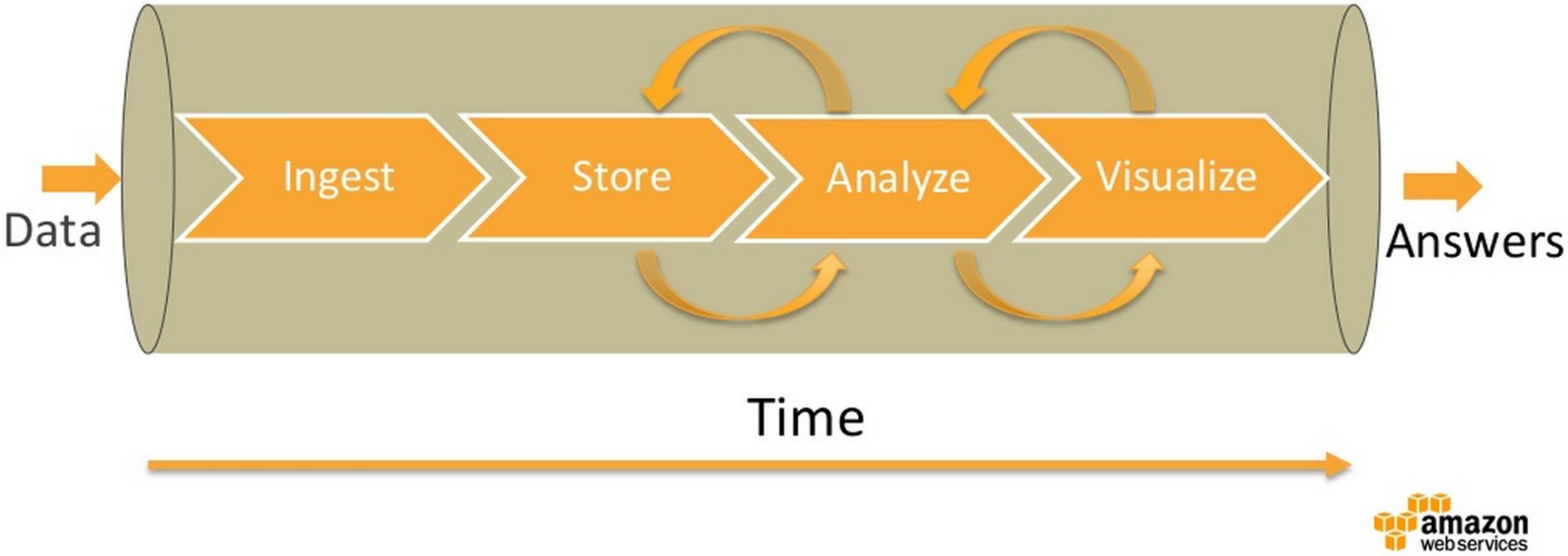
O que é Big Data?

- Conjunto de tecnologias usadas para armazenar e processar grande quantidade de dados.
- Análise de dados para diversas aplicações
  - Economia, medicina, pesquisa e desenvolvimento, etc.
- Área de grande importância
  - Crescimento de 10% ao ano; 2x maior que a da indústria de software.

Diferenciais de Big Data

- Volume: Quantidade crescente de transações e novos tipos de dados trazem problemas de armazenamento e análise de dados.
- Velocidade: Fluxos de entrada de dados cada vez maiores → necessidade de processamento mais eficaz para atender a demanda.
- Variedade: Quantidade maior de tipos de dados (principalmente de redes sociais e aparelhos móveis) dificultam a análise.

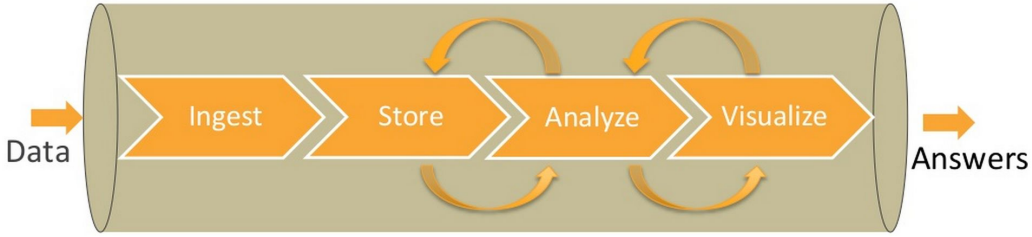
Pipeline de Big Data

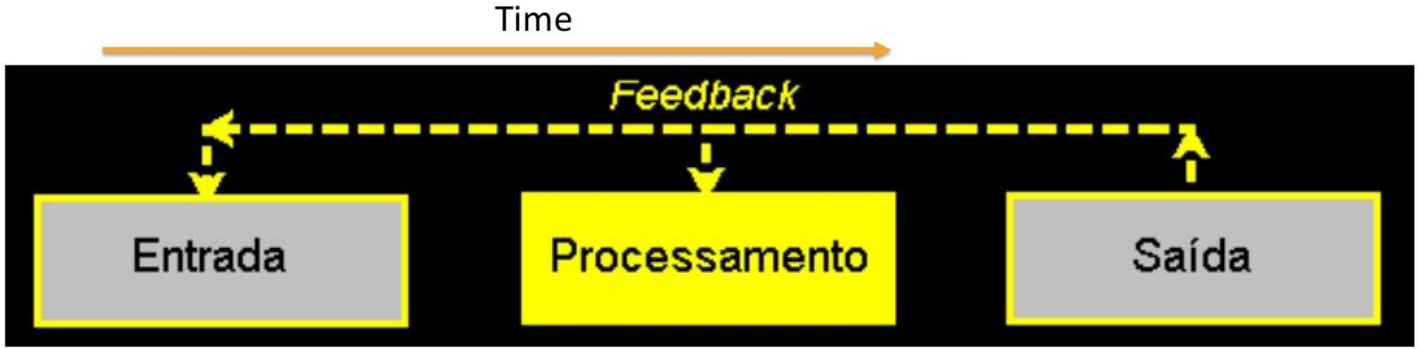


Disponível em http://pt.slideshare.net/AmazonWebServices/getting-started-with-big-data-and-hpc-in-the-cloud-august-2015

Pipeline de Big Data

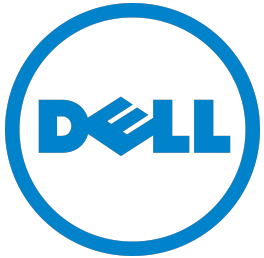
Semelhanças?

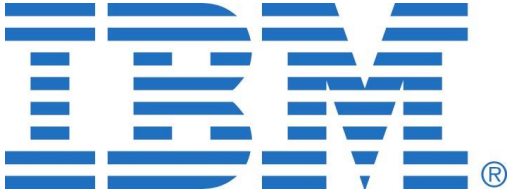




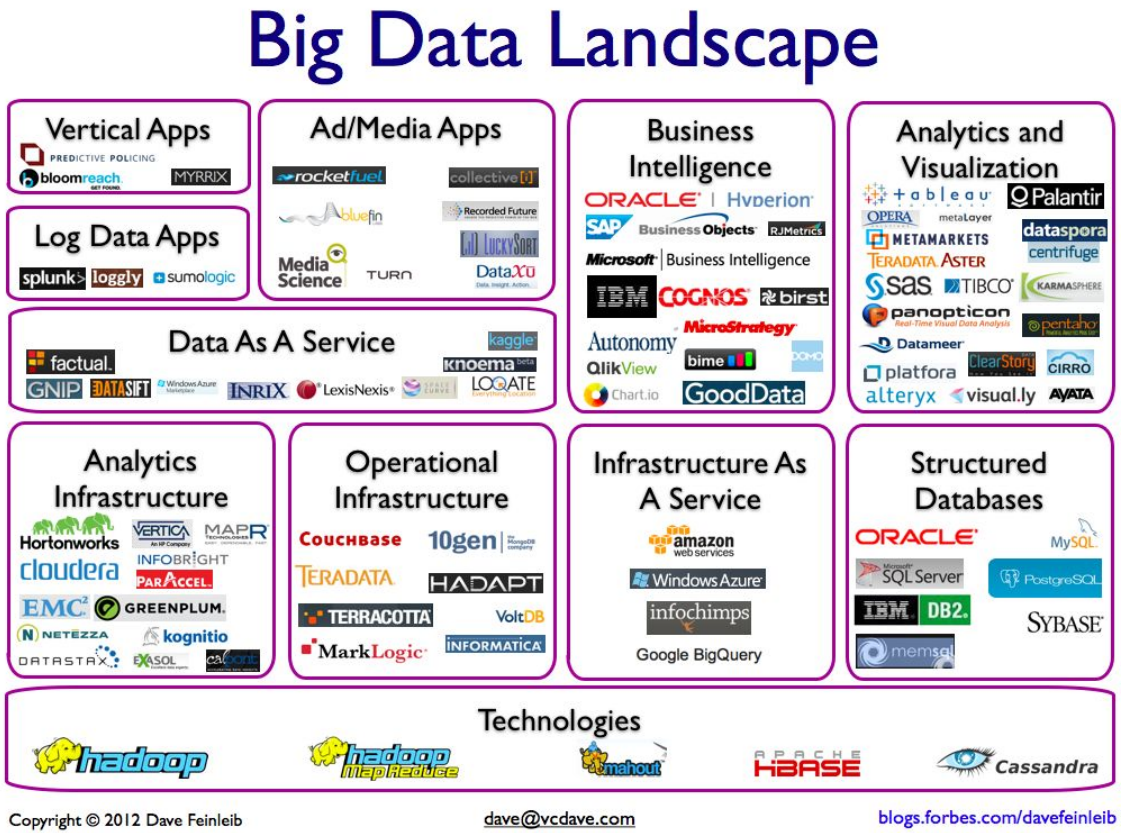
Quem trabalha com Big Data?

- Grandes investidores em Big Data (mais de 15 bilhões de dólares)





Quem trabalha com Big Data?



Disponível em http://blogs-images.forbes.com/davefeinleib/files/2014/06/big-data-landscape-jul-4-2012-00111.png

Quem utiliza esses serviços?

Cases disponíveis em https://aws.amazon.com/solutions/case-studies/all/
e https://aws.amazon.com/pt/solutions/case-studies/

Exemplo de uso de Big Data: Amazon

- Maior varejista on-line do mundo.
- Crescimento da organização → aumento no tamanho do banco de dados.

- Desafio: planejamento da utilização e custo de manutenção e *backup*.

Exemplo de uso de Big Data: Amazon

- Solução: Amazon Web Services
  - Serviço de Infrastructure as a Service (IaaS), desenvolvido para uso interno.
  - 2006: Comercialização do serviço.
  - 2010: Todos os serviços web da amazon. com migraram para o AWS.
- Atual líder no setor IaaS (10 vezes maior que seus 14 concorrentes diretos *combinados!*)

Exemplo de uso de Big Data: Philips

- Divisão de serviços de saúde da organização utililiza dados de diagnósticos e tratamentos na tomada de decisões (DSS).

- Problema: crescimento do serviço tornou sua infraestrutura incapaz de processar 37 milhões de registros
    - 434 registros por minuto - inviável!

Exemplo de uso de Big Data: Philips

- Solução: HealthSuite
  - Plataforma digital na nuvem para serviços de saúde baseada no Amazon Web Services.

- Uso de soluções de Data Warehouse para armazenamento
  - Processamento de aproximadamente 411.000 registros por minuto!

O que é Data Warehouse?

- Orientado a Assunto: dividido por áreas de negócios;
- Integrado: padroniza os dados das várias partes do sistema;
- Não Volátil: sempre inserido, nunca excluído;
- Variante no Tempo: posições históricas das atividades no tempo.

O que é Data Warehouse?

Possibilita a analise de grandes volumes de dados.

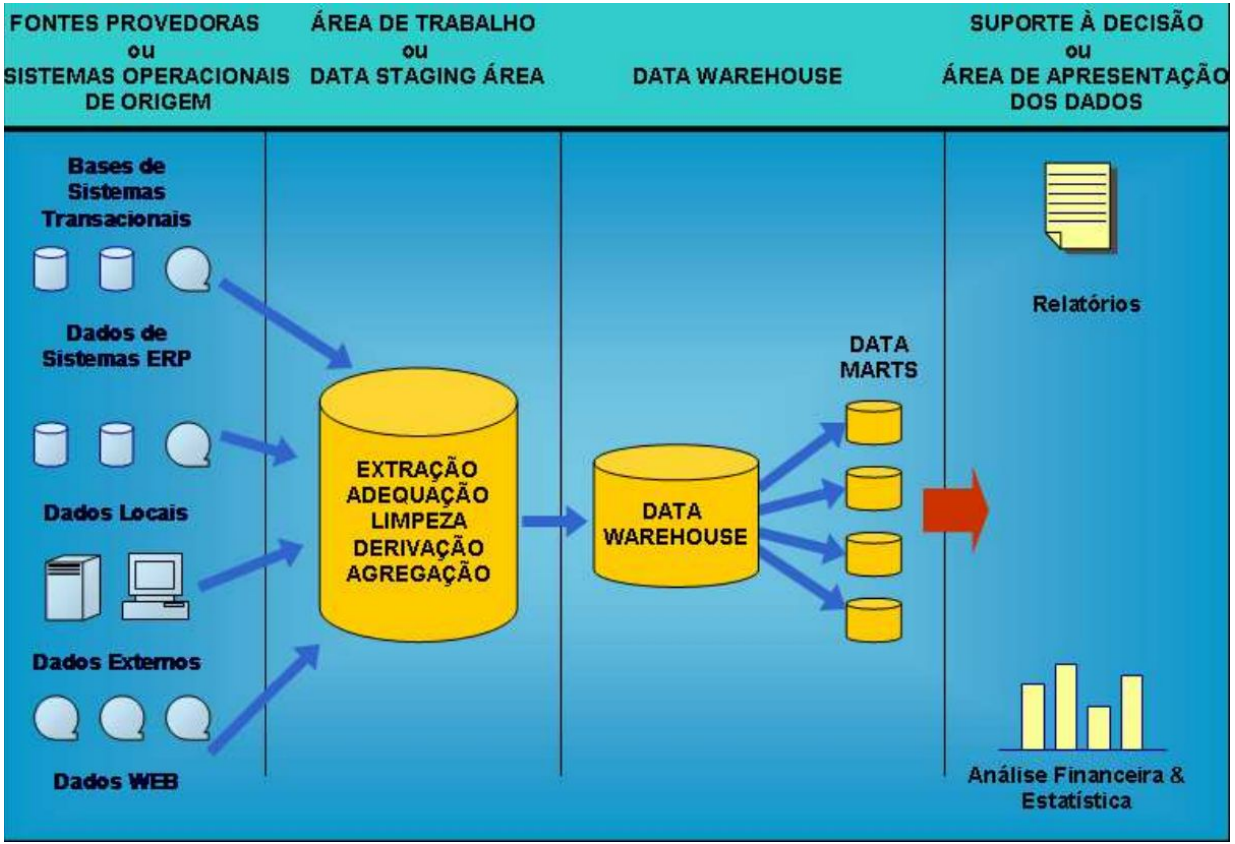
Oferece suporte a tarefa de tomada de decisão e planejamento.

O que é Data Warehouse?

"Data Warehouse é uma coleção de dados orientados por assuntos, integrados, variáveis com o tempo e não voláteis, para dar suporte ao processo de tomada de decisão."

- INMON, William H., 1990.

Elementos básicos do DW



Elementos do DW - Data Stage

Dados são extraídos geralmente de sistemas transacionais, podendo existir também dados locais e externos.

Como existem várias partes do sistema esses dados são tratados antes de serem armazenados no DW.

Elementos do DW - Armazenamento

Armazena grandes quantidades de informação.

São armazenados através de Data Marts que são pontos de acesso a subconjuntos do Data Warehouse.

Exemplo: um Data Mart financeiro poderia armazenar informações consolidadas dia-a-dia para um usuário gerencial e em periodicidades maiores (semana, mês, ano) para um usuário no nível da diretoria.

Elementos do DW – Visualização dos Dados

Os dados só são visualizados na área de consulta, ou seja, após o tratamento e armazenamento adequado.

São acessados por ferramentas de:

- OLAP (On-line Analytical Processing);
- Mineração de dados.

Elementos do DW - OLAP

São ferramentas com capacidade de análise em múltiplas perspectivas das informações armazenadas.

- Gerador de relatório;
- Visualizador de dados.

Elementos do DW - Mineração de Dados

São ferramentas com capacidade de descoberta de conhecimento relevante.

A mineração de dados não é aplicada em tempo real, é sempre sobre um conjunto de dados relacionado a um período (variáveis com o tempo).

Lideres:

- Oracle
- Teradata
- IBM
- Microsoft
- SAP

Disponível em https://www.mapr.com/sites/default/files/magic-quad-2016.jpg

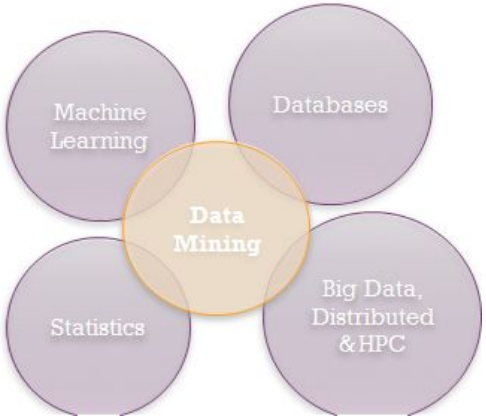
Knowledge Discovery in Databases (KDD)

- Data Mining realiza a análise de dados
- KDD: Extração de conhecimento a partir da análise da mineração de dados
- Ferramenta utilizada em diversas áreas do setor econômico
  - Marketing
  - Investimentos
  - Identificação de fraudes financeiras
  - Telecomunicações
  - ...

O que é Data Mining?

- Se você procurar na internet?
  - “Databases”... “Big Data”
  - “Machine Learning”...”Statistics”
  - Jargões de marketing?

- Disciplinas Relacionadas: Estatistica, Machine Learning, Big Data
- Machine Learning aplicada.



O que é Data Mining?

- Procura de padrões nos dados.
- Solucionar problemas através da analise de dados já presentes no banco de dados.
- Extrair conhecimento dos dados.

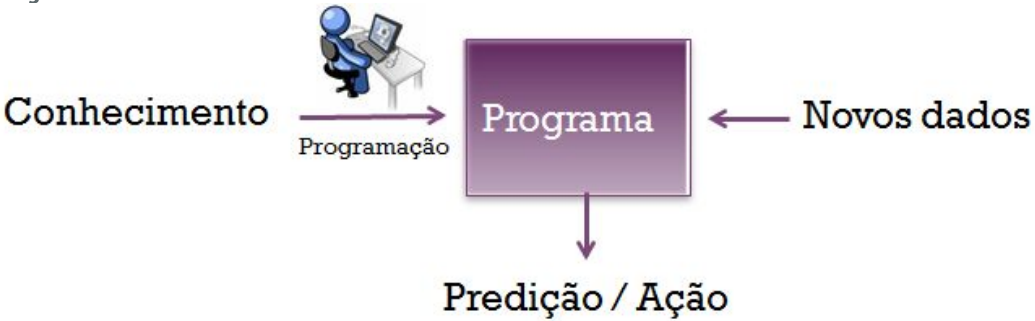
O que é Data Mining?

- Computação Tradicional

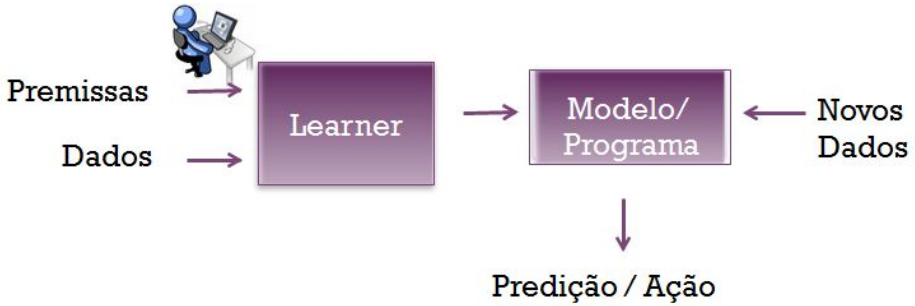
- Machine Learning

O que é Data Mining?

- Computação Tradicional



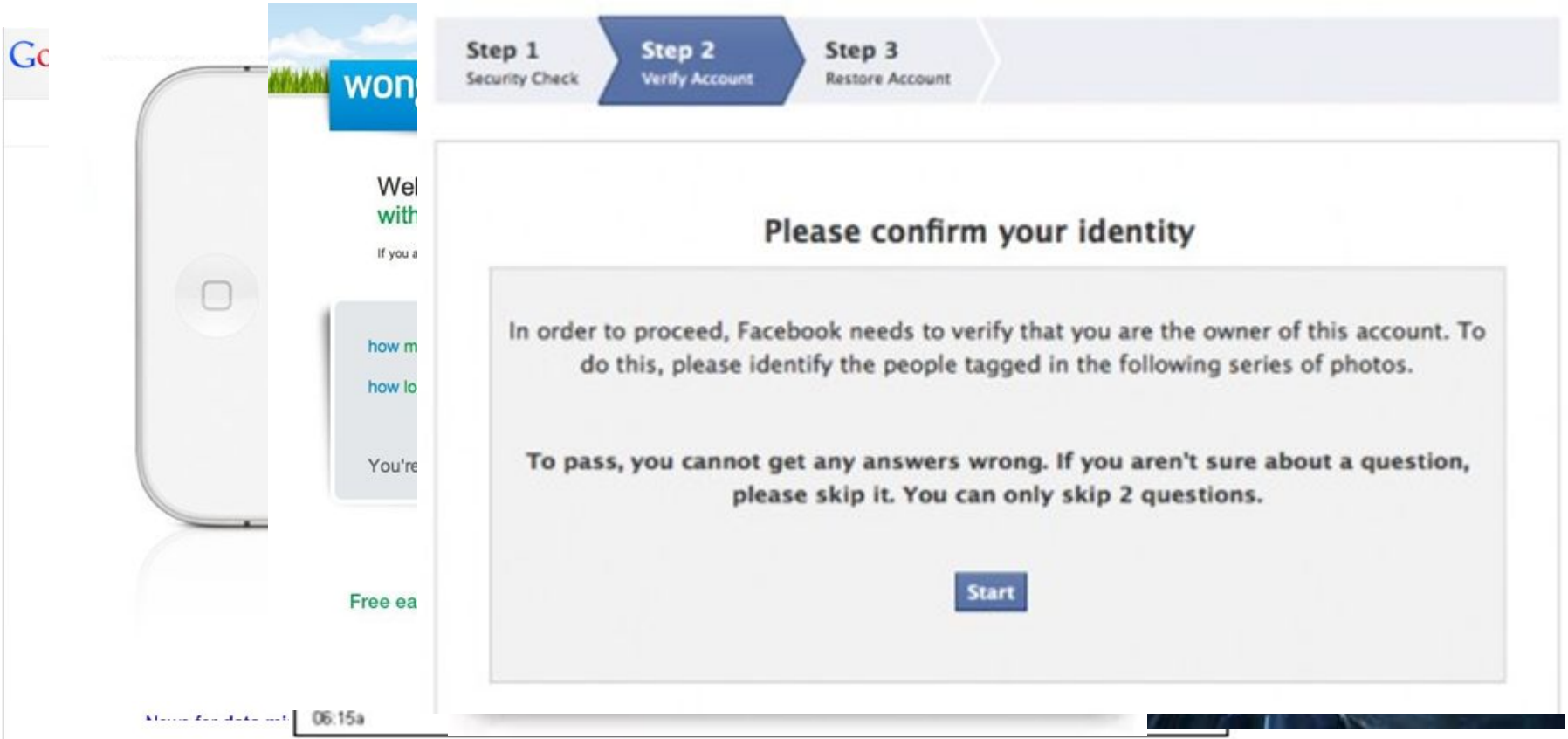
- Machine Learning



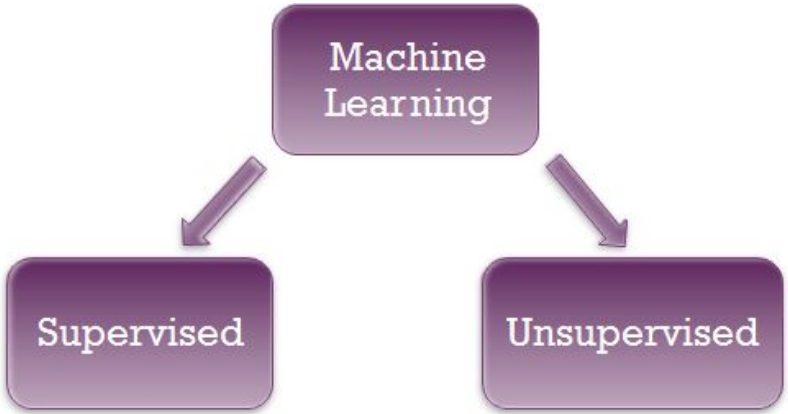
Por que fazer Data Mining?

- Um grande volume de dados é produzido.
  - Fontes vistas no começo da apresentação.
- Um recurso valioso a ser explorado.
- Dados em si são inúteis: são necessários métodos para extrair informação automaticamente deles.
  - Informações extraidas são os padrões ocultos nos dados.

Exemplos de Machine Learning



Uma taxonomia



Alvo

Table 1.3 Weather Data with Some Numeric Attributes

Outlook	Temperature	Humidity	Windy	Play
Sunny	85	85	false	no
Sunny	80	90	true	no
Overcast	83	86	false	yes
Rainy	70	96	false	yes
Rainy	68	80	false	yes
Rainy	65	70	true	no
Overcast	64	65	true	yes
Sunny	72	95	false	no
Sunny	69	70	false	yes
Rainy	75	80	false	yes
Sunny	75	70	true	yes
Overcast	72	90	true	yes
Overcast	81	75	false	yes
Rainy	71	91	true	no

Supervised:

- Atributo alvo.
- Encontrar uma regra para predizer valor do atributo alvo.

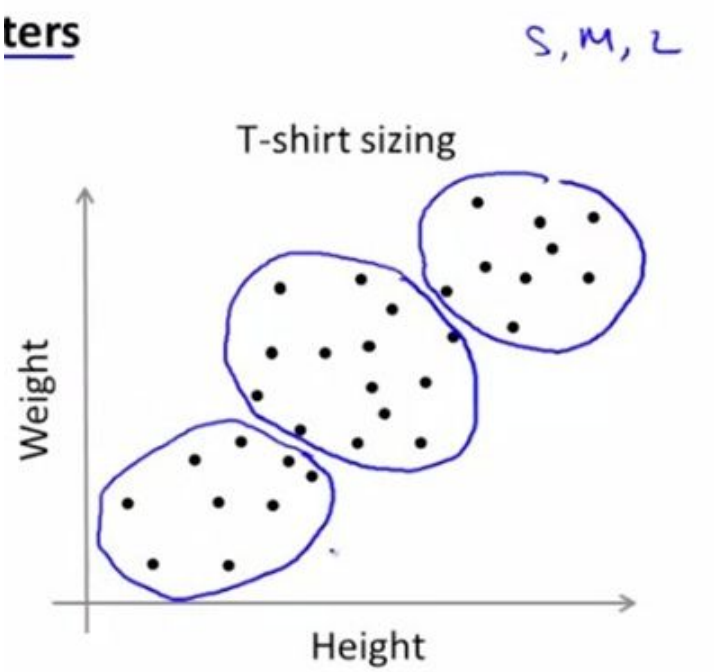
Unsupervised:

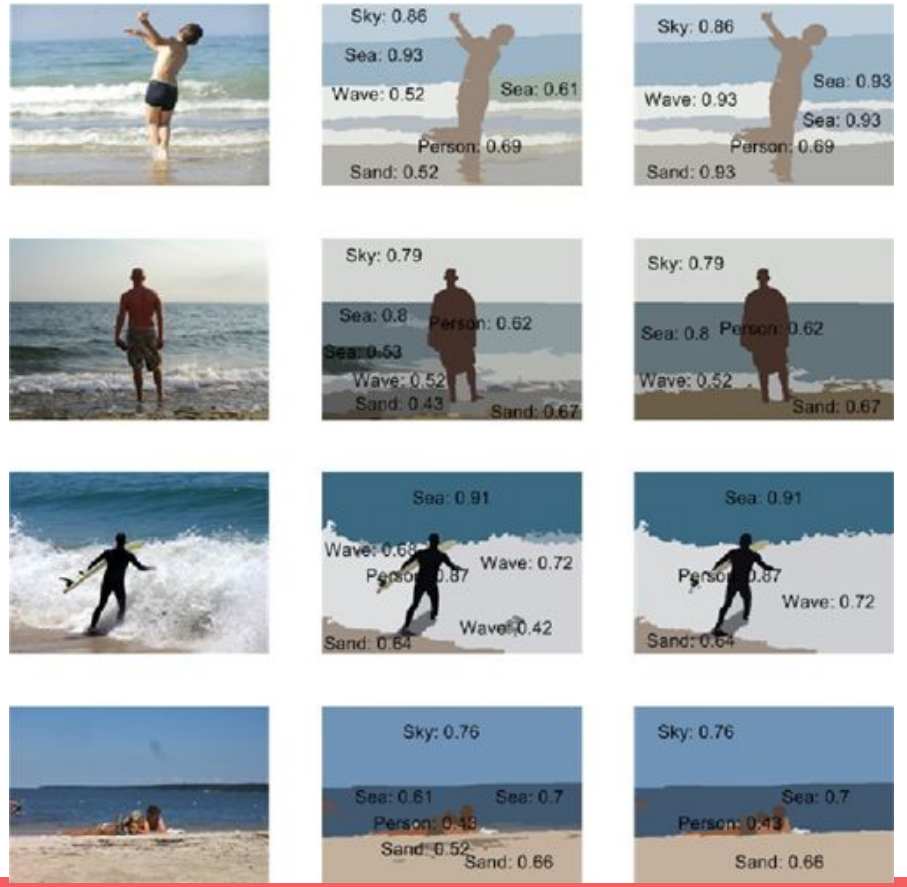
- Sem alvo.
- Entender, resumir, encontrar padrões, explicar os dados.

Unsupervised Machine Learning

Exemplos:

- Processamento de imagens.
- Perfil de consumidores.

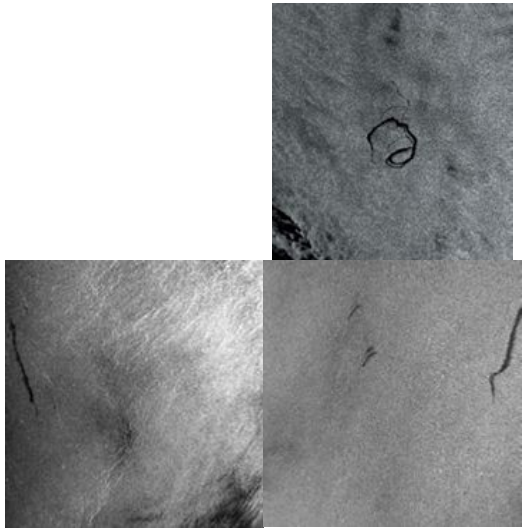




Data Mining em Negócios

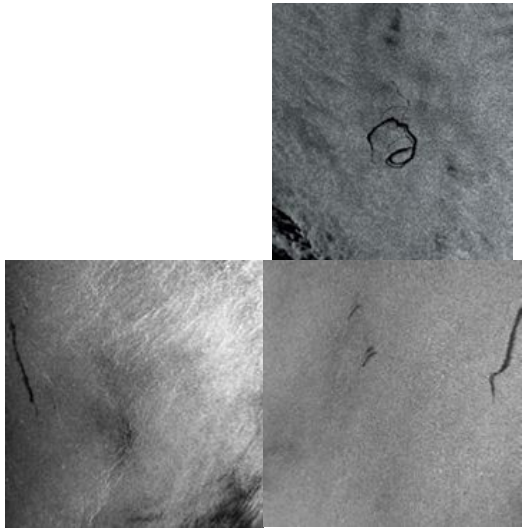
Data Mining: Coleta de Dados

- As vezes você recebe uma base de dados com atributos relevantes.
  - Normalmente você precisa decidir que dados irá coletar e armazenar na base de dados.
- Exemplo: Detectar vazamentos de óleo em imagens de satélite.
  - Vazamentos de óleo são manchas negras com tamanho e formato variado.
  - Regiões parecidas podem ser causadas pelo vento.
- Decisões de coleta de dados
  - Qual resolução para o sensor do satélite?
  - Eu preciso de informações sobre o vento?
- Decisões de coleta tem impacto em:
  - Custo de aquisição
  - Desempenho do sistema



Data Mining: Pré-processamento dos dados

- Exemplo: Detectar vazamentos de óleo em imagens de satélite.
  - Vazamentos de óleo são manchas negras com tamanho e formato variado.
  - Regiões parecidas podem ser causadas pelo vento.
- Difícil ter consistência com imagens. Extração de atributos:
  - Tamanho da região.
  - Formato da região.
  - Intensidade.
  - Quantidade de reentrâncias.
  - Proximidade de outras regiões
- Não é apenas o que coletar.
  - Mas como codificar os dados.



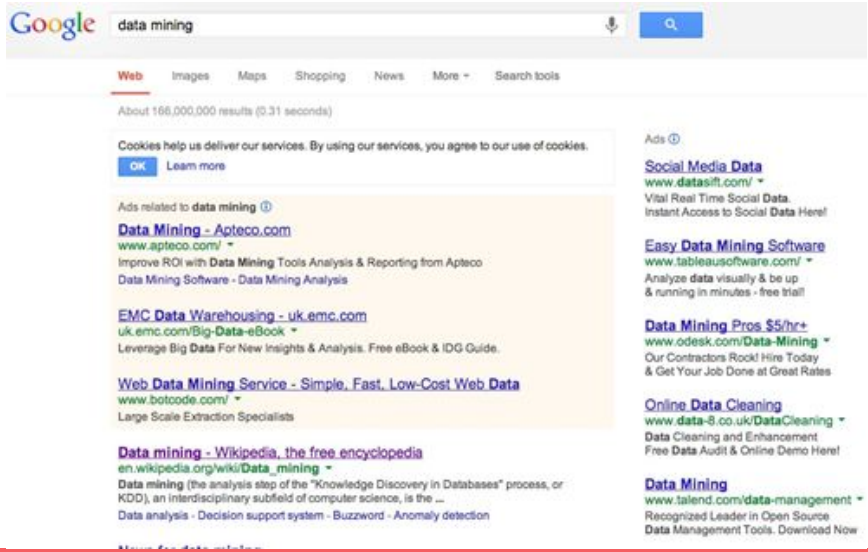
Data Mining em Negócios

Data Mining: Avaliação

- Avaliação e testes de softwares convencionais.
  - Ele produz a saída correta?
  - Qual é o custo computacional e de memória?
- Machine Learning tem uma abordagem estatística.
  - Você quase nunca tem um procedimento “correto”.
  - … Apenas um procedimento melhor ou pior (taxa de erro).
- Avaliação e testes de operações de Data Mining
  - Qualidade da saída?
    - Quão consistente é a saída?
    - Quão ruim são os erros?
  - Qual é o custo computacional e de memória?
  - Qual é a relação entre qualidade da saída e custo computacional?

Data Mining: Exemplo de Negócio

- Valor para o negócio?
  - Mostrar propagandas relevantes.
  - Pay per click
- Dados de entrada + saída?
  - Termos da pesquisa
  - Clique?
- Modelo?
  - Prever propagandas atraentes.
- Dados adicionais?
  - Perfil do usuário.
- Como avaliar?
  - Taxa de cliques (click through rate)
- Impacto da computação?
  - Custos.
  - Experiência do usuário.



Exemplo de uso Data Mining - WalMart

- Baixo nível de estoque e ressuprimento constante
  - Prevê cada item por loja
  - Identificação de padrão de consumo por loja
- Exemplo Clássico:
  - Vendas de Fralda / Cerveja

Exemplo de uso Data Mining - Itau

- Envio de mala direta para correntistas
  - Menos de 2% dos correntistas respondiam as promoções.
  - Gasto desnecessário com o serviço de correios.
- Com a análise de dados
  - Taxa de respostas as promoções subiram para 30%.
  - Gasto com o serviço de correios reduzido para um quinto.

Exemplo de uso Data Mining - ShopKo

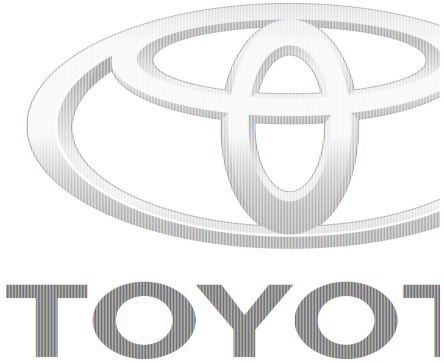
- Sofria com a concorrência da WalMart

- Venda de produtos através da venda de outros produtos
  - Resistiu a concorrência em 90% dos mercados
  - Aumentou as vendas

Exemplo de uso Data Mining - Sprint

- Prevê se o consumidor deseja trocar de companhia telefonica no periodo de 2 meses
  - 61% de segurança na previsão
  - Uso de marketing para não perder o cliente
  - Evitou a perda de 120.000 clientes e 35 milhões de dólares

Exemplo de uso Data Warehouse - Toyota



- Toyota Motor Sales Usa - Distribuidora de veiculos
- No final dos anos 90 enfrentou dificuldades com:
  - problemas na cadeia de fornecimento
  - custo de armazenagem elevado
  - atraso na entrega para os revendedores
- Toneladas de dados e relatórios sem direcionamento
- Configurado um sistema para fornecer dados precisos em tempo real
- Em 2000 utilização de um Data Warehouse da Oracle e configurado um novo sistema.
- Em poucos dias:
  - Descobriu-se que a Toyota era cobrada duas veze por envio especial de trem (erro de US$ 800.000)
  - Aumento do volume de carros negociados em 40%
  - Tempo de trânsito reduzido em mais de 5%

Exemplo de uso Data Warehouse - Vivo

- 6 empresas de telefonia trasacionavam cerca de 2 bilhões de dados diariamente
- Trabalhavam com ferramenta de Business Intelligence e processos distintos
- Tempo de resposta muito alto
- Integração de todos os sistema em um único Data Warehouse
- Economia de US$ 28 milhões de dólares
- Marketing melhor dirigido
- Otimização do uso da rede e identificação de falhas

Perguntas?

Referências

- JAMES, Josh. “How Much Data Is Created Every Minute?”. Junho 2012. Disponível em <https://www.domo.com/blog/2012/06/how-much-data-is-created-every-minute/> . Acesso em 17/05/2016.
- JAMES, Josh. “Data Never Sleeps 2.0”. Abril 2014. Disponível em <https://www.domo.com/blog/2014/04/data-never-sleeps-2-0/> . Acesso em 17/05/2016.
- CHAKRABARTI, S et al. “Data Mining Curriculum: a Proposal”. Abril 2006. Disponível em <http://www.kdd.org/curriculum/view/introduction> . Acesso em 17/05/2016.
- FAYYAD, Usama; PIEATETSKY-SHAPIRO, Gregory; SYMTH, Padhraic. “From Data Mining to Knowledge Discovery in Databases”. 1996. Disponível em <http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>. Acesso em 17/05/2016.

Referências

- Gartner IT Glossary. “What is Big Data?” 2012. Dispoível em <http://www.gartner.com/it-glossary/big-data/>. Acesso em 17/05/2016.
- Gartner. “Gartner Says Solving ‘Big Data’ Challenges Involves More Than Just Managing Volumes Of Data”. Disponível em <http://www.gartner.com/newsroom/id/1731916>. Acesso em 17/05/2016.
- OLIVEIRA, Marcell. Data Warehouse. Abril 2008. Disponível em <http://www.datawarehouse.inf.br/Academicos/A%20PUBLICAR_DATA_WAREHOUSE_MARCELL_OLIVEIRA.pdf>. Acesso em 22/05/2016.
- FEINLEIB, Dave. “The Big Data Landscape”. Junho 2012. Disponível em <http://www.forbes.com/sites/davefeinleib/2012/06/19/the-big-data-landscape/#10c6861d3b8a>. Acesso em 17/05/2016.

Referências

- DIONÍZIO, Leonardo. Data Warehouse. Disponível em <http://www.devmedia.com.br/data-warehouse/12609>. Acesso em 23/05/2016.
- Amazon Web Services. “AWS Case Study: Philips uses Amazon Redshift for Large Data Workloads”. Disponível em <https://aws.amazon.com/solutions/case-studies/philips-redshift/>. Acesso em 23/05/2016.
- LINOFF, Gordon S.; BERRY, Michael J. A. Data Mining Techniques: For Marketing, Sales and Customer Relationship Management. Third Edition 2011. 888p.
- JULIANELLI, Leonardo. “Big Data: Como lidar com a diversidade de formatos?”. Disponivel em <http://www.ilos.com.br/web/tag/big-data/> . Acesso em 23/05/2016.

Referências

- MATTISON , Rob. “Selecting Tools for Data Mining”. Disponivel em <http://www.uniforum.org/publications/ufm/sept96/mining.html> . Acesso em 23/05/2016.
- HOOFNAGLE, Chris J; KING, Jennifer. “Comsumer Information Sharing: Where the Sun Still Don’t Shine”. 2007. Disponível em <http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1137990>. Acesso em 23/05/2016.