PROGRAMA DE
PÓS-GRADUAÇÃO EM
ENGENHARIA DE SISTEMAS E PRODUTOS

ADS
IFBA

Big Data

Renato Novais
Departamento de Computação
renato@ifba.edu.br

Qual o tamanho?

72 horas
De vídeos por minuto no YouTube[1]

145 bilhões
emails enviados diariamente[1]

40TB
dados produzidos por um experimento do LHC/Cern em um segundo[1]

40 mil
Consultas no Google por segundo[2]

3.5 bi
Consultas no Google por dia[2]

[1]http://marciaconner.com/blog/data-on-big-data/, jul. 2012.
[2]http://www.internetlivestats.com/google-search-statistics/, jun 2018

facebook

1.45 bi

de usuários ativos (em média) diariamente[1]

2.2 bilhões

de usuários por mês[1]

83 milhões

perfis falsos[1]

300 mi

de uploads de fotos por dia[1]

[1]https://zephoria.com/top-15-valuable-facebook-statistics/, jun. 2018.

100 milhões
Ligações por voz diariamente[1]

1.5 bilhão
de usuários [1]

55 milhões
Ligações por vídeo diariamente[1]

1 bilhão
de usuários ativos diariamente[1]

65 bilhões
mensagens enviadas diariamente[1]

[1]https://expandedramblings.com/index.php/whatsapp-statistics/, jun. 2018.

117 milhões
de assinaturas[1]

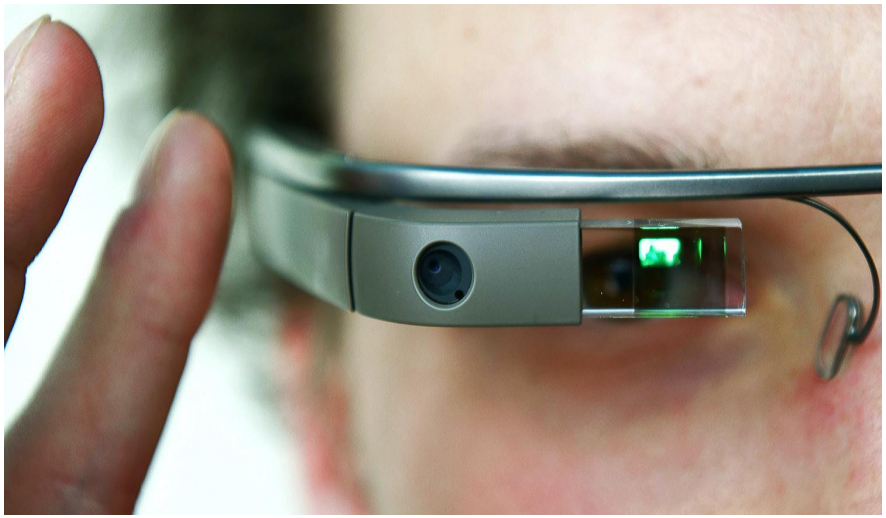
1 bilhão
de horas de vídeo assistidas por semana[1]

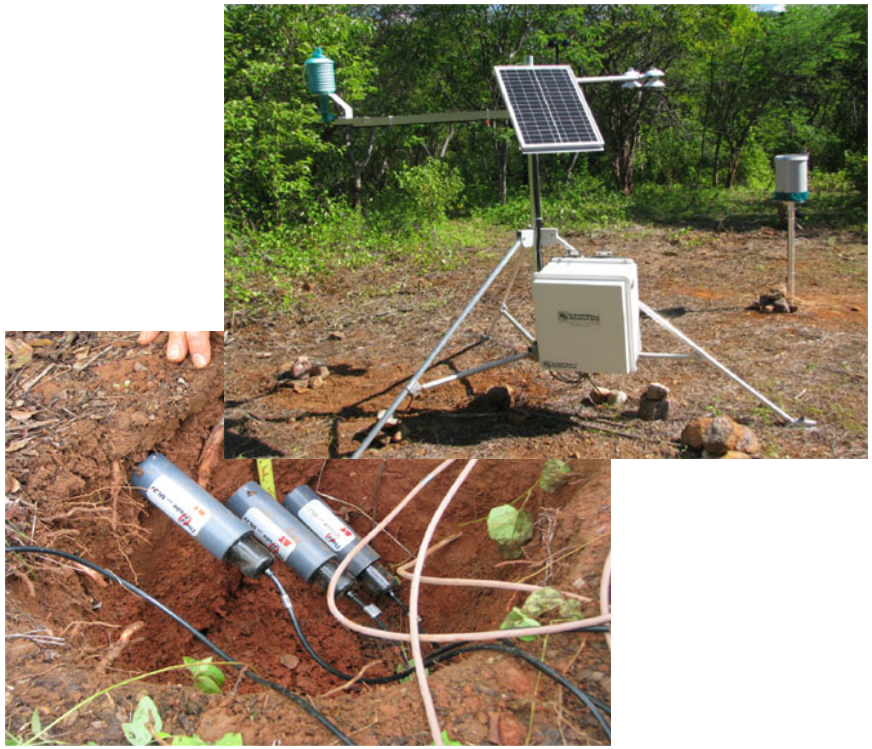
15 bilhões
receita prevista para 2018[1]

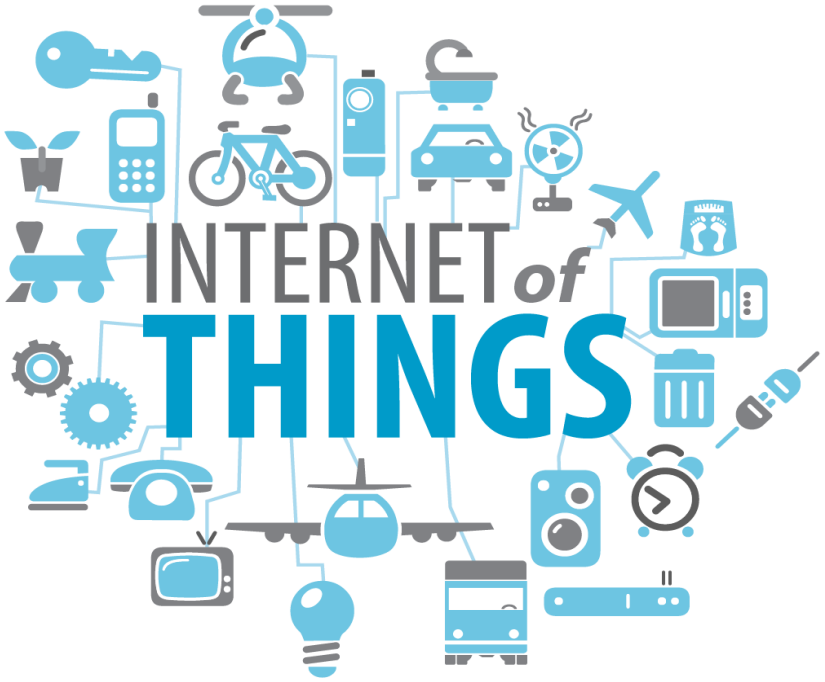
6 bilhões
gastos para produzir conteúdo em 2017[1]

[1]https://expandedramblings.com/index.php/netflix_statistics-facts/, jun. 2018.

Qual a origem?






INTERNET of
THINGS

Existem mais dispositivos conectados à internet que seres humanos na Terra

Estimasse 20 bilhões de dispositivos para 2020

https://www.youtube.com/watch?v=1SNoNTaWFIo

Qual o problema?

90%

produzidos nos
dois últimos anos

4.4ZB tamanho do
universo digital

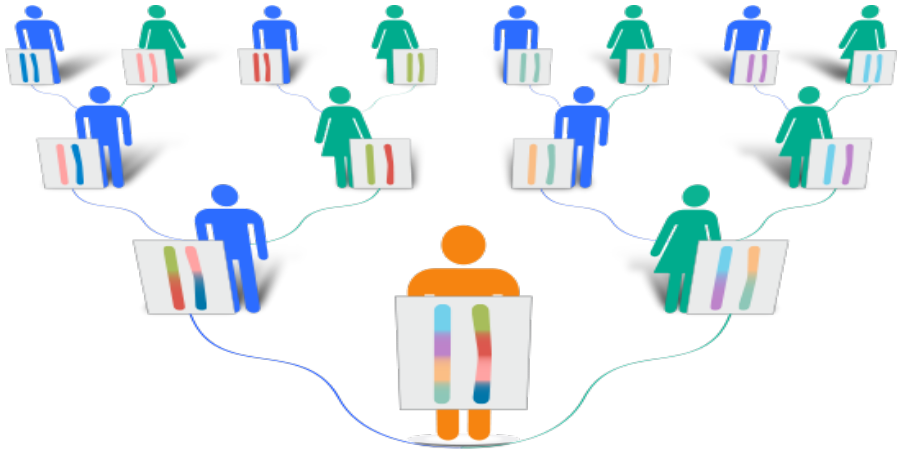
1ZB = 1024EB
1EB = 1024PB
1PB = 1024TB
1TB = 1024GB

1% analisado para extrair
alguma informação

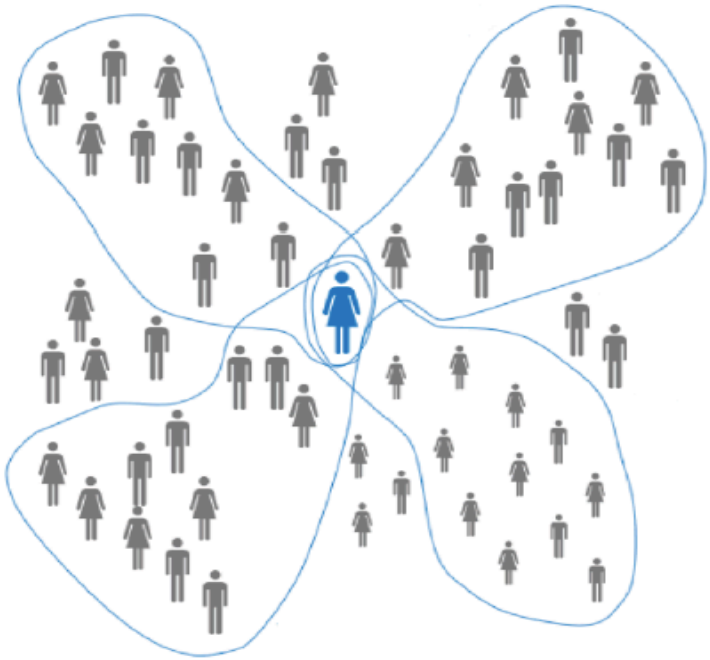
33% armazenados

http://www.emc.com/leadership/digital-universe/2012iview/big-data-2020.htm, dez. 2012.

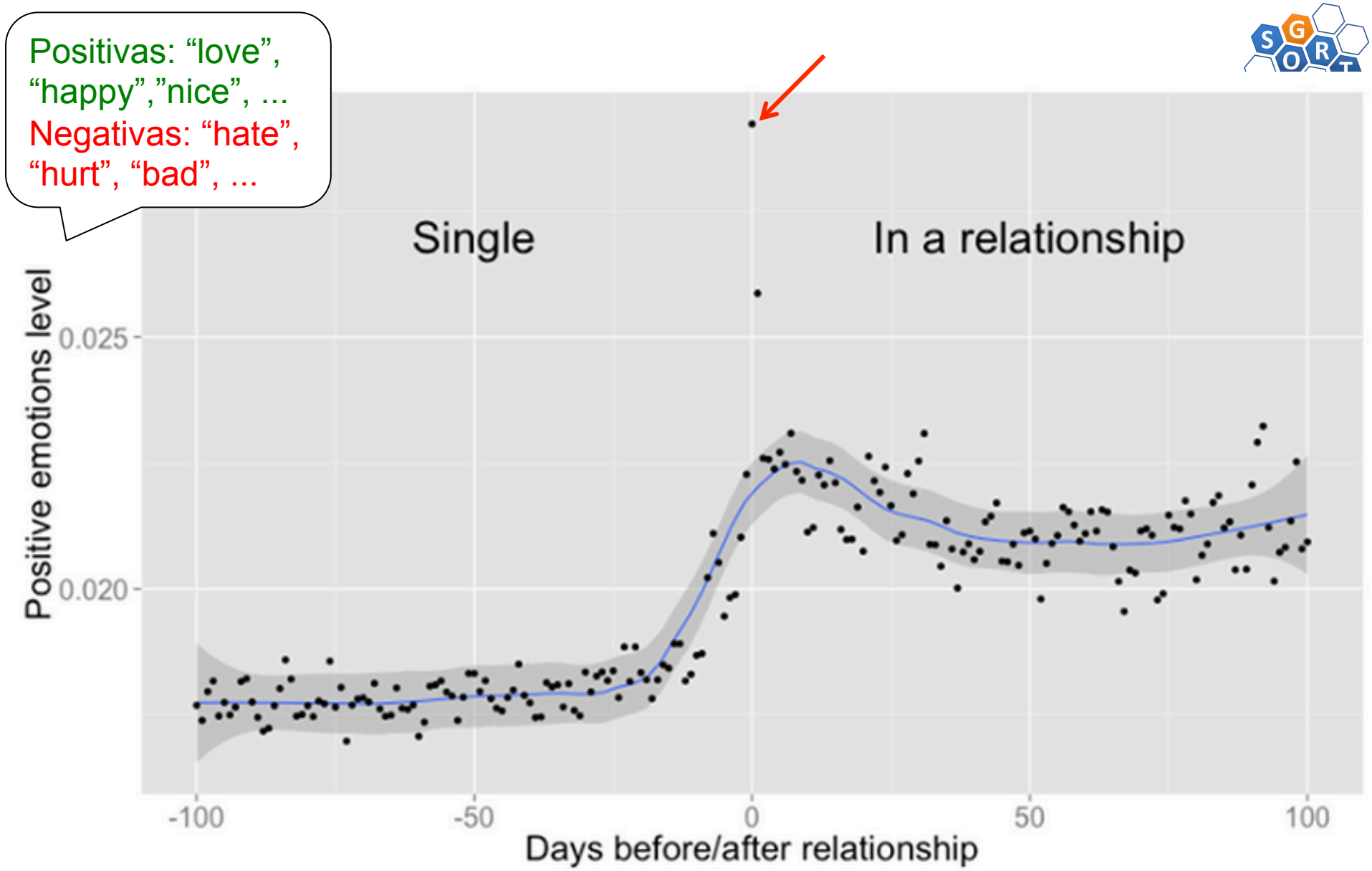
O que poderíamos encontrar?



SETI HOME

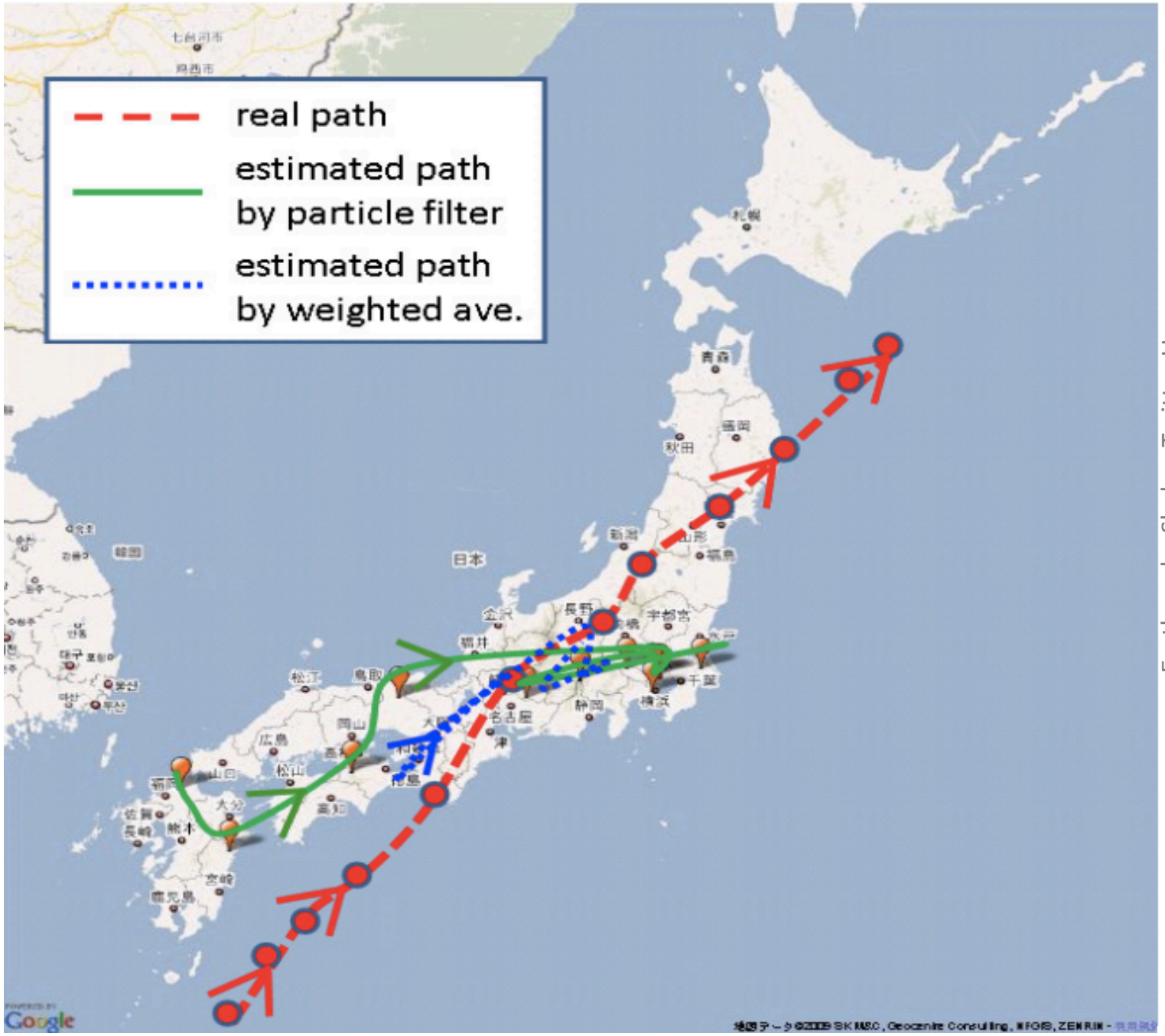


As pessoas ficam mais positivas após um relacionamento?



https://www.facebook.com/notes/facebook-data-science/the-formation-of-love/10152064609253859, fev. 2014.

Podemos utilizar o Twitter para prever eventos naturais?



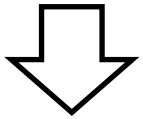
Earthquake Shakes Twitter Users:
Real-time Event Detection by Social Sensors, WWW, 2010.

fonte: http://akrayasolutions.com/big-data/

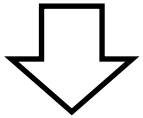
"Grande quantidade de dados que não consegue ser tratada com a tecnologia atual"

"Ferramentas e tecnologias utilizadas para extrair informação de uma grande quantidade de dados"

Aumento do volume de dados

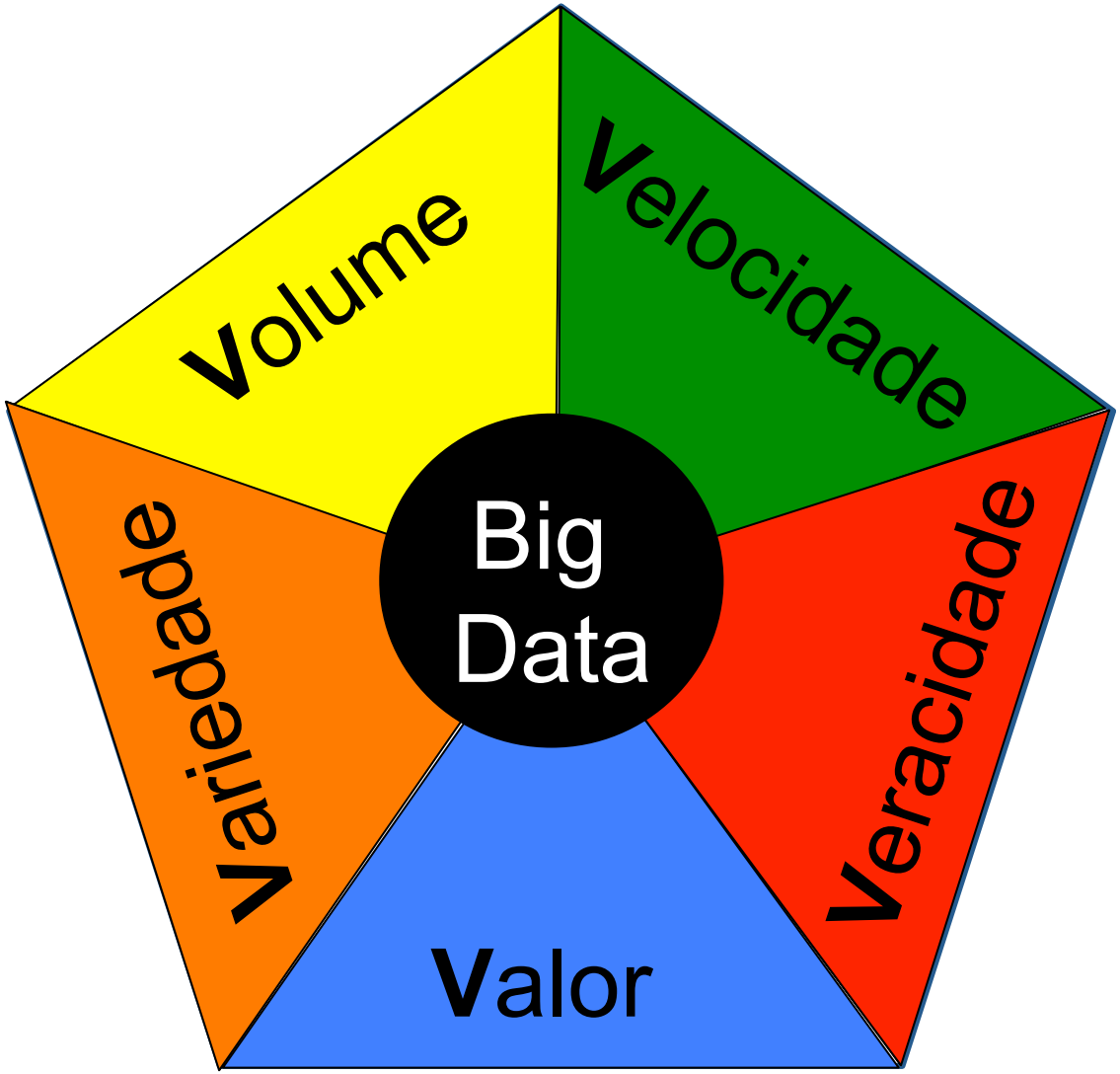


Limitação dos bancos de dados relacionais

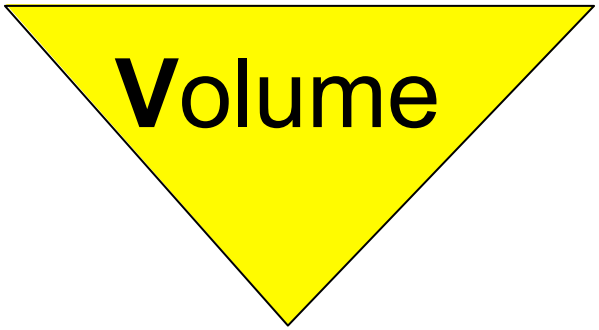


Uso de computação paralela com dispositivos de baixo custo

Propriedades dos dados (5 Vs)

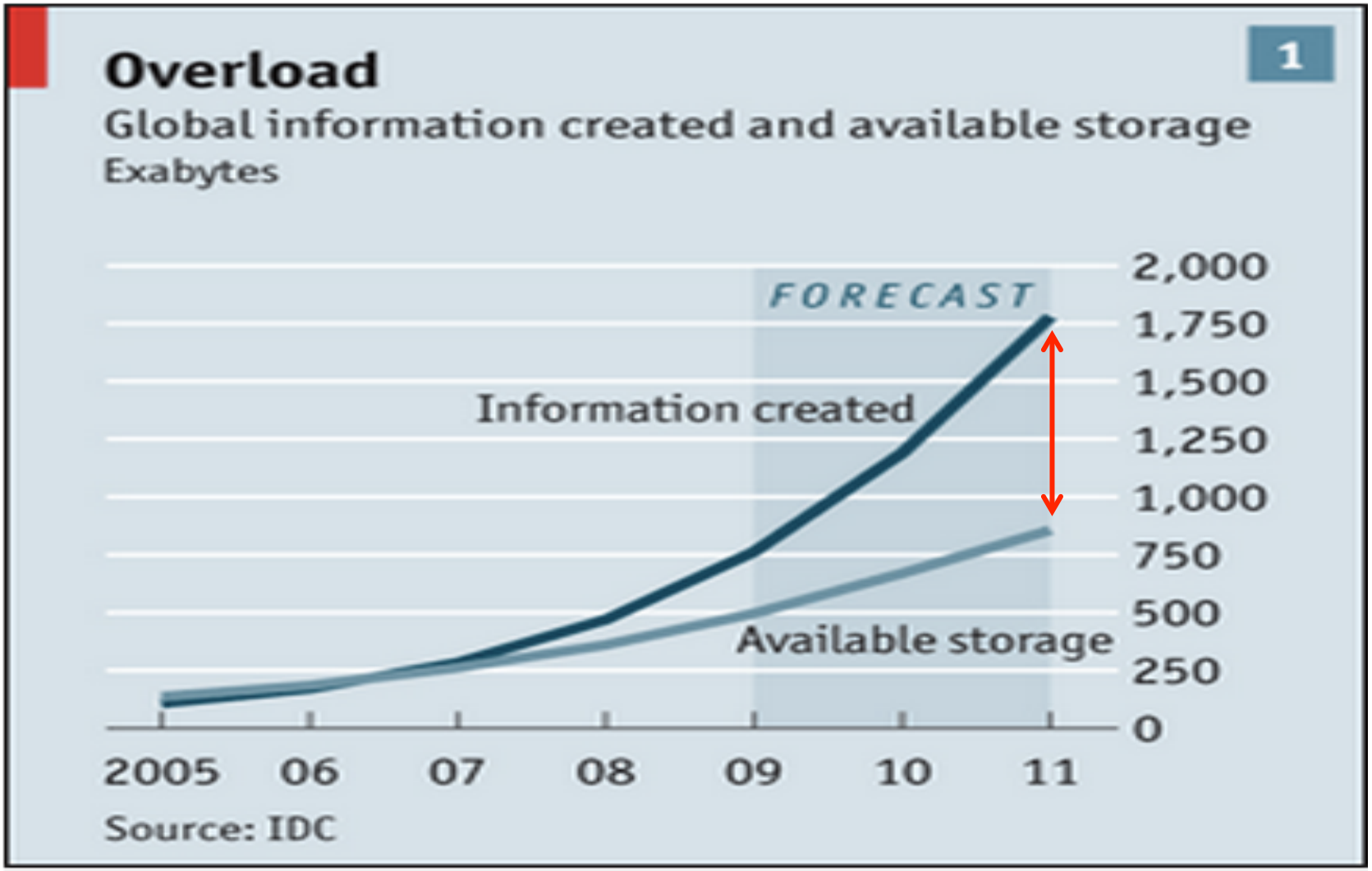


Volume



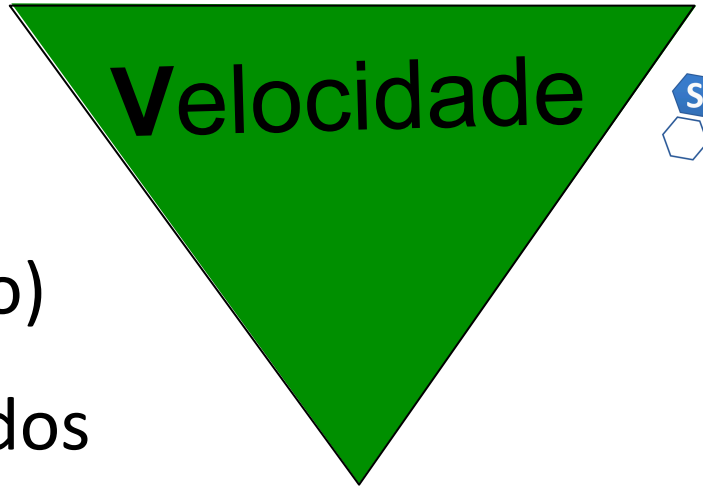
- Crescimento exponencial
- 3x mais dados transmitidos que armazenados
- PCs e servidores armazenam maior parte dos dados
- A maior parte dos dados é produzida por humanos
- **Tendência**
  - mais dados armazenados em dispositivos móveis
  - mais dados gerados pela Web das coisas (não humanos)

Volume



http://www.economist.com/node/15557443

Velocidade



- Dados em movimento (fluxo)
- Dados precisam ser analisados rapidamente (próximo do tempo real)
  - não dá tempo extrair, transformar e carregar esses dados em outro formato
- **Tendência**
  - processar parte dos dados (soluções aproximadas)
  - armazenar os dados em memória principal
  - consultas analíticas em tempo real

Variedade

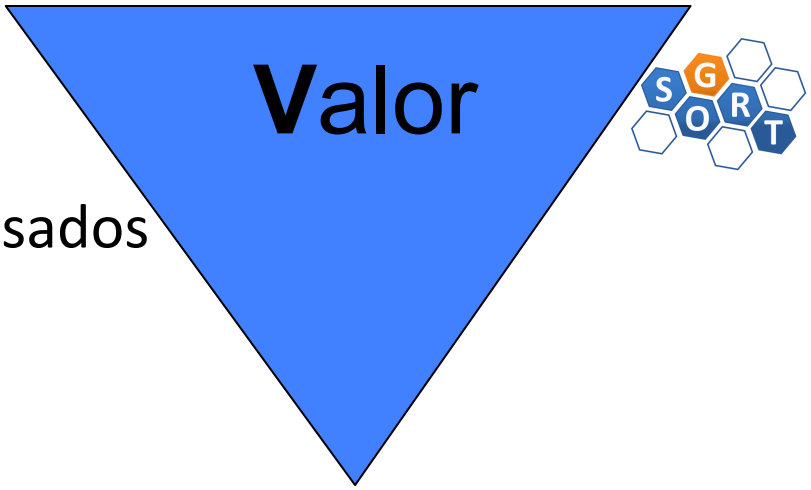
- Dados estruturados
  - tabelas
- Semi-estruturados
  - XML
- Não estruturados
  - sensores, clickstream, logs, textos, áudio, vídeo, imagens, ...
- Hoje, apenas 10% dos dados são estruturados
- **Tendência**
  - crescimento maior dos **dados estruturados**
  - consultas avançadas em dados não estruturados
  - consultas analíticas diretamente sobre dados estruturados

Veracidade

- 1 em cada 3 analistas de negócio não acreditam na informação que eles usam para tomar decisões
- Dados incompletos, errados ou desatualizados
- Dados de diferentes fontes com valores diferentes
- **Tendência**
  - dados probabilísticos
  - consultas aproximadas

http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

Valor



- Dados só têm valor quando processados e analisados
- Alguns setores
  - **Varejista**: produto certo para o comprador certo
  - **Financeiro**: prevenir fraudes em tempo real
  - **Provedores de conteúdo**: sistemas de recomendação
  - **Cidades**: congestionamento, segurança, ...
  - **Utilidade**: melhor utilização de recursos hídricos e elétricos
- **Tendências**
  - novas formas de visualizar os dados
  - consultas próximas do tempo real

Intel. Turn Big Data into Big Value, 2013.

7 insights em Bigdata (#1)

- Os dados invadiram toda a indústria e negócio e são agora um importante fator de produção
  - Nós estamos gerando tanto dado que é impossível de armazenar fisicamente
  - Big data está em qualquer setor da economia

Quais seriam exemplos de Bigdata?

Trabalho 1

- Apresentar um exemplo interessante de aplicação de big data em algum setor importante da sociedade no Brasil ou no Mundo (setor público, saúde, educação, agronegócio, gerenciamento de emergência, etc)
- Explicar quais são os 5 Vs do exemplo excolhido

7 insights em Bigdata (#2)

- Bigdata cria valor em todos os sentidos
  - **Criar Transparência**: tornar os dados mais facilmente acessíveis para diferentes stakeholders em tempo hábil pode trazer um grande valor.
    - Abrir os dados do serviço público. Seja para o público externo ou para os diferentes setores da instituição.
  - **Permitir experimentação para descobrir necessidades, expor variabilidade e melhorar performance**: a medida que você armazena mais e mais dados, você pode ter um conhecimento mais preciso e detalhado sobre tudo da sua empresa.

7 insights em Bigdata (#2)

- Bigdata cria valor em todos os sentidos
  - **Segmentar a população e customizar as ações**: personalizar os produtos e serviços
    - melhorar publicidade, atender diferente pessoas com necessidades diferentes.
  - **Apoiar a tomada de decisão humana com algoritmos automatizados**: Análises sofisticadas podem melhorar substancialmente a tomada de decisões, minimizar os riscos e revelar *insights* valiosos que, de outra forma, permaneceriam ocultos.

7 insights em Bigdata (#2)

- Bigdata cria valor em todos os sentidos
  - **Inovando novos modelos de negócios, produtos e serviços**: O Big Data permite que as empresas criem novos produtos e serviços, aprimorem os já existentes e inventem modelos de negócios totalmente novos.

7 insights em Bigdata (#3)

- O uso de Big data será um ponto-chave de concorrência e crescimento para empresas individuais
  - Está se tornando um dos principais meios para as empresas líderes superarem seus pares
  - Ajudará a criar novas oportunidades de crescimento e novas categorias de empresas, como as que agregam e analisam dados do setor.

7 insights em Bigdata (#4)

- O uso de big data submeterá novas ondas de crescimento de produtividade e de excedente
  - permitir que as organizações façam mais com menos e produzam resultados de maior qualidade
  - Os dados podem até ser aproveitados para melhorar os produtos à medida que são usados
    - Produtos que aprendem os hábitos e necessidades de seu usuário

7 insights em Bigdata (#5)

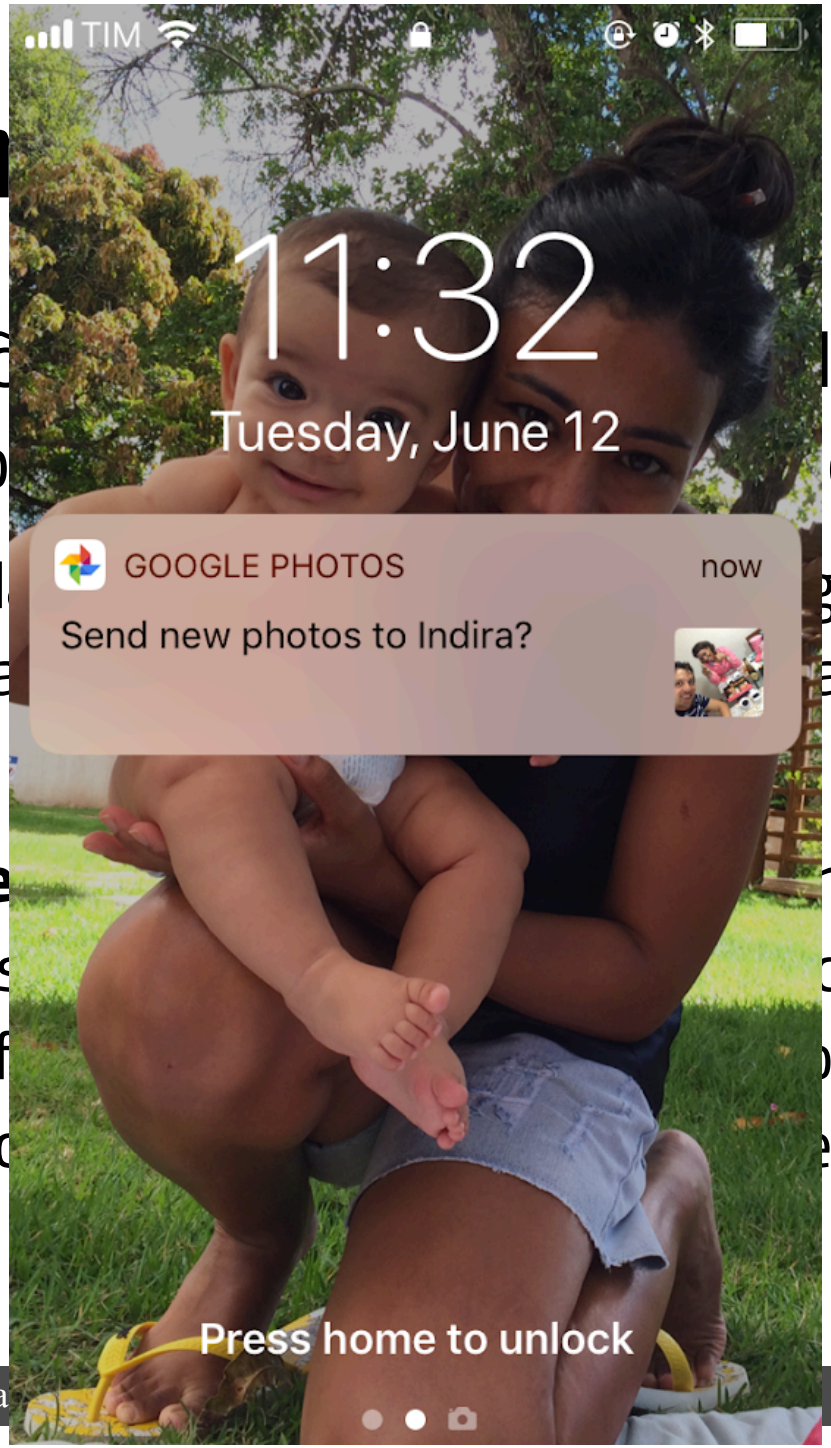
- O uso de big data é relevante para todos os setores, mas alguns setores terão ganhos maiores
  - Setores que naturalmente produzem dados irão sair na frente
  - TI é transversal.

7 insights em Bigdata (#6)

- Haverá uma falta de talento necessário para as organizações tentarem a vantagem dos grandes dados
  - Particularmente pessoas com conhecimento avançado em estatística e *marchine lerning* (...e análise de dados em geral)

7 insights em

- Várias questõ las para capturar o po dados
  - **Política de d** gitalizados, logo é preciso cuida ança, propriedade intelectual e
  - **Tecnologia e** r de big data, as organizações cnologias (por exemplo, sof mputação e análise) e téc e análises).



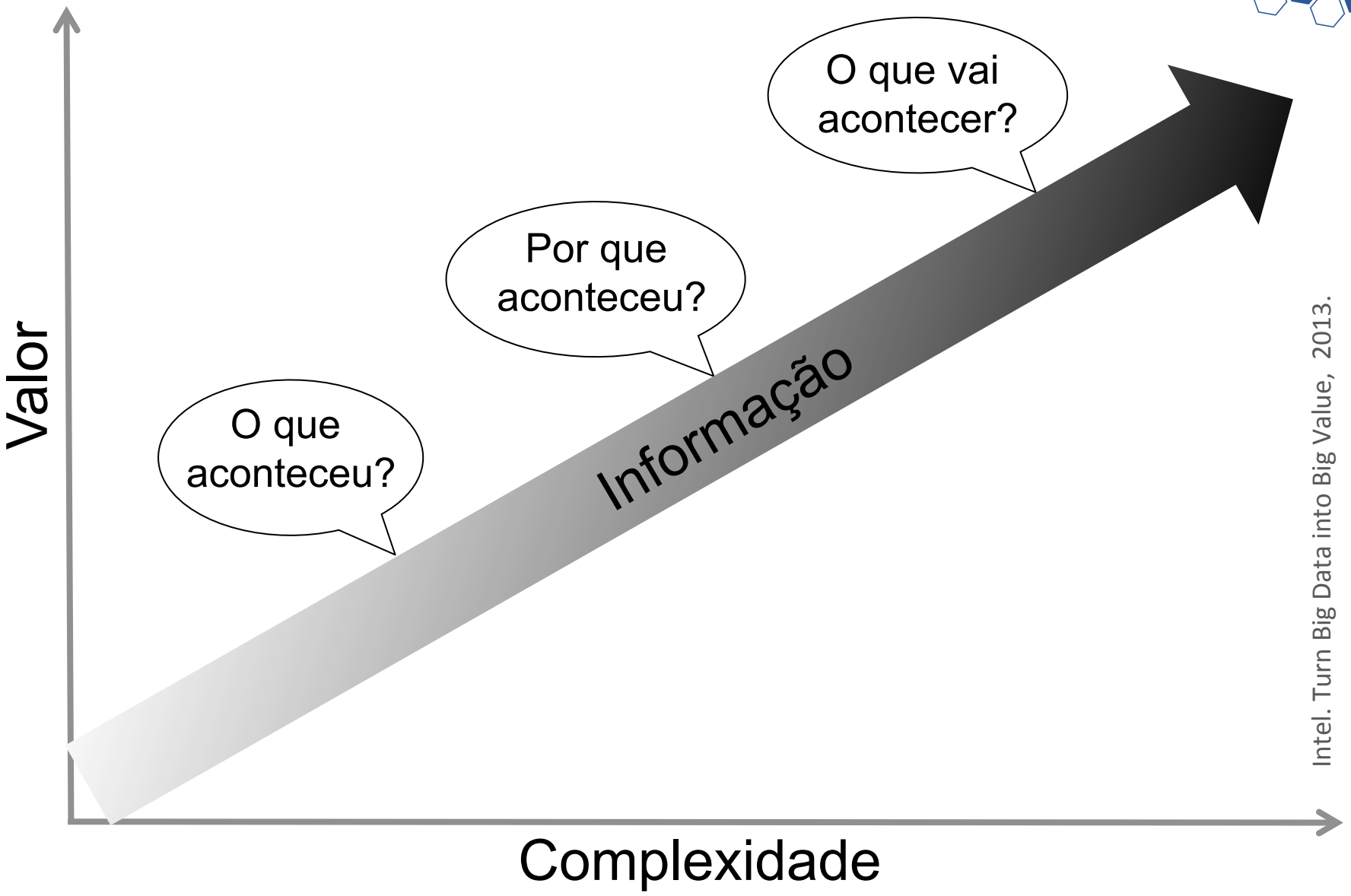
7 insights em Bigdata (#7)

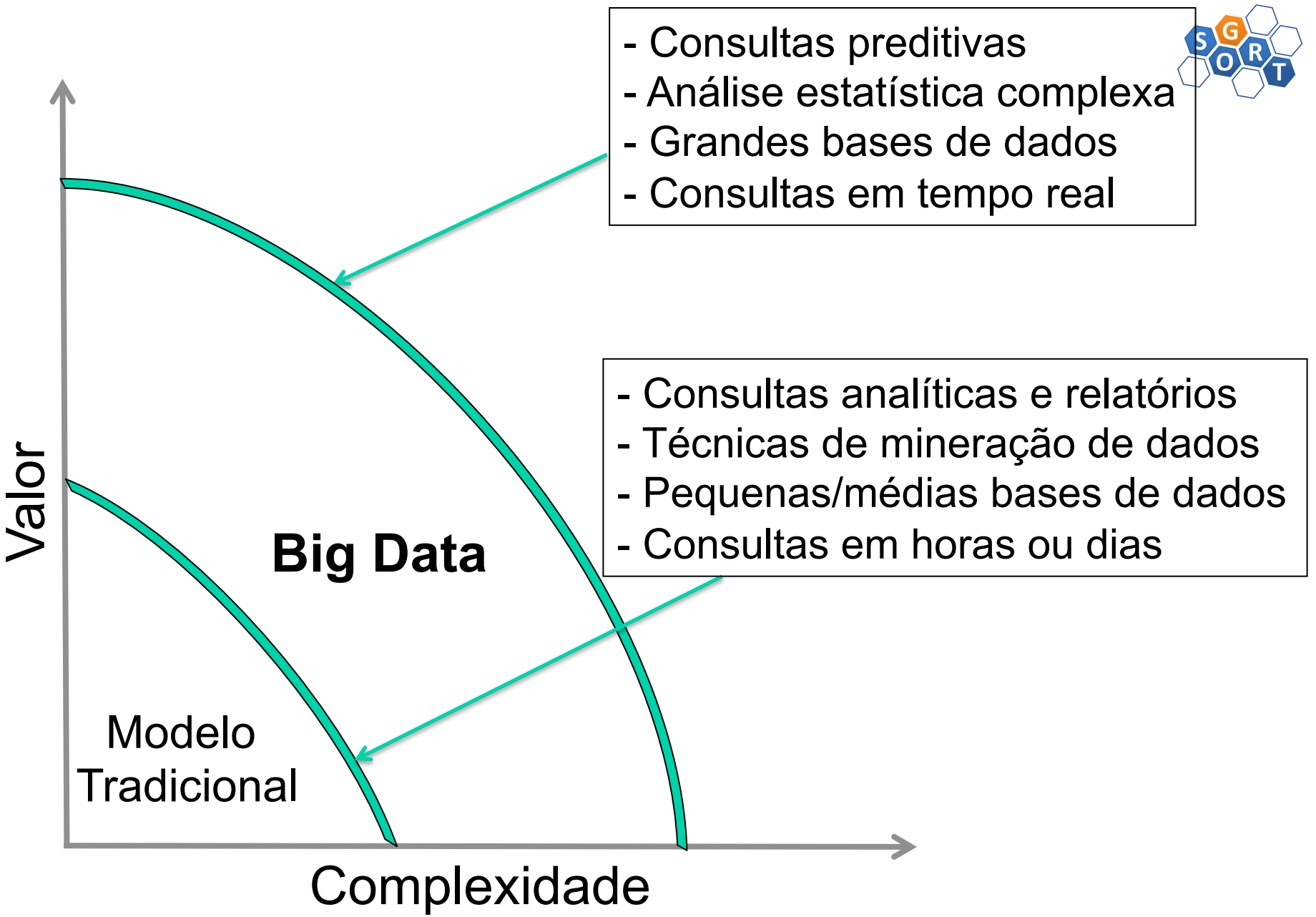
- Várias questões terão que ser apontadas para capturar o potencial total dos grandes dados
  - **Mudança organizacional e talento:** Os líderes organizacionais muitas vezes não entendem o valor do big data e também como desbloquear esse valor.
  - **Acesso aos dados:** Para possibilitar oportunidades transformadoras, as empresas precisarão cada vez mais integrar informações de várias fontes de dados.

7 insights em Bigdata (#7)

- Várias questões terão que ser apontadas para capturar o potencial total dos grandes dados
  - **Estrutura da indústria:** Setores com uma relativa falta de intensidade competitiva e transparência de desempenho, junto com indústrias nas quais os pools de lucros são altamente concentrados, tendem a ser lentos para aproveitar totalmente os benefícios do Big Data.
    - Os líderes de organizações e formuladores de políticas terão de considerar como as estruturas do setor poderiam evoluir em um grande mundo de dados se quiserem determinar como otimizar a criação de valor no nível de firmas, setores e economias individuais como um todo.

O que está guiando o desenvolvimento em Big Data?


Valor
O que aconteceu?
Por que aconteceu?
O que vai acontecer?
Informação
Complexidade
Intel. Turn Big Data into Big Value, 2013.


- Consultas preditivas
- Análise estatística complexa
- Grandes bases de dados
- Consultas em tempo real
- Consultas analíticas e relatórios
- Técnicas de mineração de dados
- Pequenas/médias bases de dados
- Consultas em horas ou dias
Valor
Big Data
Modelo
Tradicional
Complexidade

Modelo tradicional

Análise estruturada e repetitiva

O usuário determina as perguntas

TI estrutura os dados para responder as perguntas

Big Data

Análise iterativa e exploratória

TI prepara plataforma para consultas exploratórias e criativas

O usuário explora questões que podem ser perguntadas. Cria e processa novas perguntas.

Martin Pavlik, IBM Big Data Plataform Overview, 2013.

Quais técnicas?

Técnicas para lidar com Big Data

- Há muitas técnicas para lidar com big data, dentre elas:
  - A/B Testing
  - Association Rule learning
  - Classification
  - Cluster Analysis
  - Crowdsourcing
  - Data fusion and Data Integration
  - Data Mining
  - Ensemble Learning
  - Genetic Algorithmic
  - Machine Learning
  - Natural Language Processing
  - Neural Networks
  - Network analysis
  - Optimization
  - Pattern recognition
  - Predictive Modeling
  - Regression
  - Sentiment Analysis
  - Signal Processing
  - Spatial analysis
  - Statistics
  - Supervised Learning
  - Time Series Analysis
  - Unsupervised Learning
  - Visualization

etc...

Trabalho 2

- Escolher 1 técnica para fazer uma apresentação, contendo
  - Teoria (definições)
  - Aplicações possíveis
  - Um exemplo prático aplicado a análise de dados
  - Tempo de apresentação: 1 aula (140 min)
    - Mostrar teoria
    - Fazer atividades práticas com a turma

A/B Testing

- Uma técnica em que um grupo de controle é comparado com uma variedade de grupos de teste, a fim de determinar que tratamentos (i.e., alterações) melhoraria uma dada variável objetiva, (i.e., Taxa de resposta de marketing).
  - Exemplo: utilizar o teste A/B para identificar alterações nas páginas web que podem ser negativas ou positivas
  - Grupo A (controle): versão atual do site
  - Grupo B (tratamento): nova versão do site

A/B Testing

- Conversion -> converter a visita ao objetivo do site (vendas, leitores, etc)
- Medir o desempenho de uma variação (A ou B) significa medir a taxa na qual ela converte os visitantes em realizadores de metas.

https://vwo.com/ab-testing/, jun. 2018.

A/B Testing

- Big Data permite que um grande número de testes seja executado e analisado, garantindo que os grupos tenham tamanho suficiente para detectar diferenças significativas (ou seja, estatisticamente significativas) entre os grupos de controle e tratamento.

Association Rule Learning

- Um conjunto de técnicas para descobrir relações interessantes, ou seja, “regras de associação”, entre variáveis em grandes bancos de dados.
- Essas técnicas consistem em uma variedade de algoritmos para gerar e testar possíveis regras.
- Aplicação
  - a análise de cesta de compras, na qual um varejista pode determinar quais produtos são frequentemente comprados juntos e usar essa informação para marketing. Ex: compradores de supermercado que compram fraldas também tendem a comprar cerveja).

Association Rule Learning

- As regras de Associação têm como premissa básica encontrar elementos que implicam na presença de outros elementos em uma mesma transação.

Exemplo de base de dados com 4 itens e 5 transações.

transação	leite	pão	manteiga	cerveja
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

https://pt.wikipedia.org/wiki/Regras_de_associa%C3%A7%C3%A3o

Association Rule Learning

- Métricas para identificar associação
  - O **suporte** sup(X) de um conjunto X é definido como a proporção de transações da base de dados que contém esse conjunto.
  - A **confiança** de uma regra é definida

Exemplo de base de dados com 4 itens e 5 transações.

transação	leite	pão	manteiga	cerveja
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

$$\mathrm{conf}(X \Rightarrow Y) = \mathrm{supp}(X \cup Y)/\mathrm{supp}(X).$$

Por exemplo, a regra {leite, pão} => {manteiga} tem uma confiança de 0.2/0.4 = 0.5 na base de dados, o que significa que para 50% das transações que contém leite e pão a regra está correta.

https://pt.wikipedia.org/wiki/Regras_de_associa%C3%A7%C3%A3o

Association Rule Learning

- Métricas para identificar associação
  - O **lift** de uma regra é definido como

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(Y) \times \text{supp}(X)}$$

A regra {leite, pão} => {manteiga} possui um lift de 0.2/(0.4 x 0.4) = 1.25.

  - Interpretação
    - Lift (X => Y) > 1 se X e Y têm correlação positiva
    - Lift (X => Y) ≈ 1 se X e Y são independentes
    - Lift (X => Y) < 1 se X e Y têm correlação negativa

Exemplo de base de dados com 4 itens e 5 transações.

transação	leite	pão	manteiga	cerveja
1	1	1	0	0
2	0	1	1	0
3	0	0	0	1
4	1	1	1	0
5	0	1	0	0

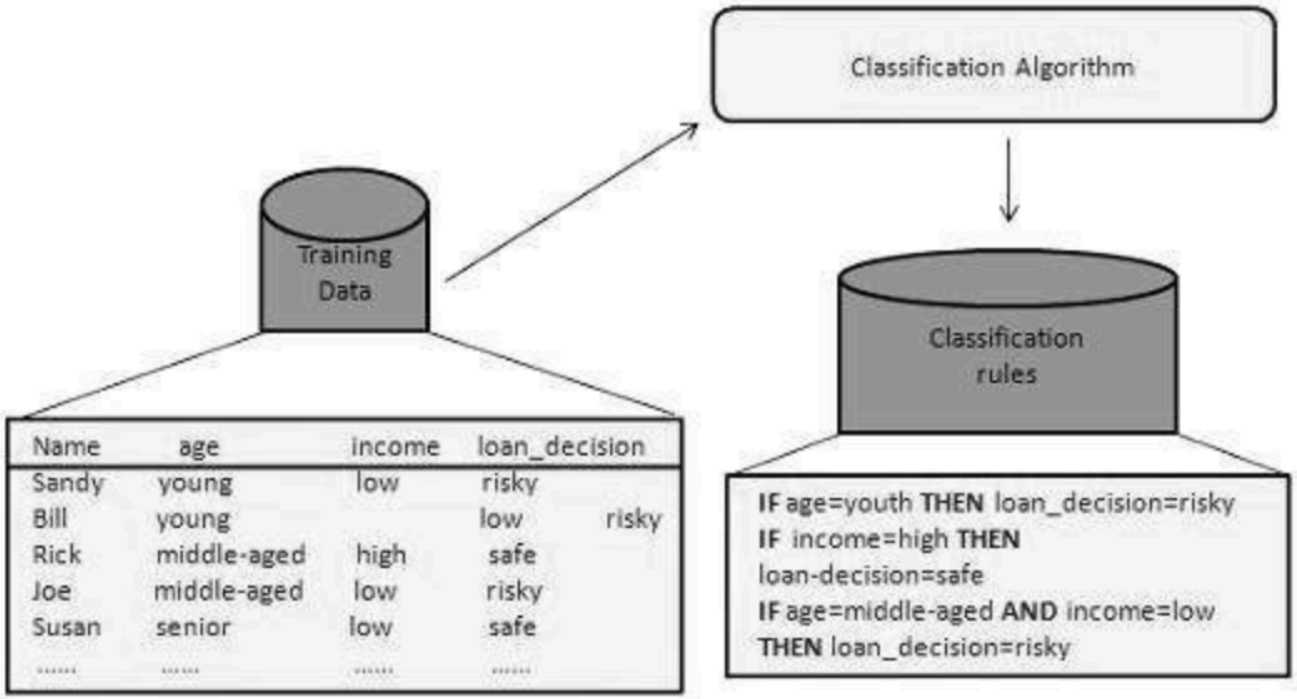
https://www.coursera.org/learn/process-mining/lecture/fk3JX/1-6-association-rule-learning

Classification

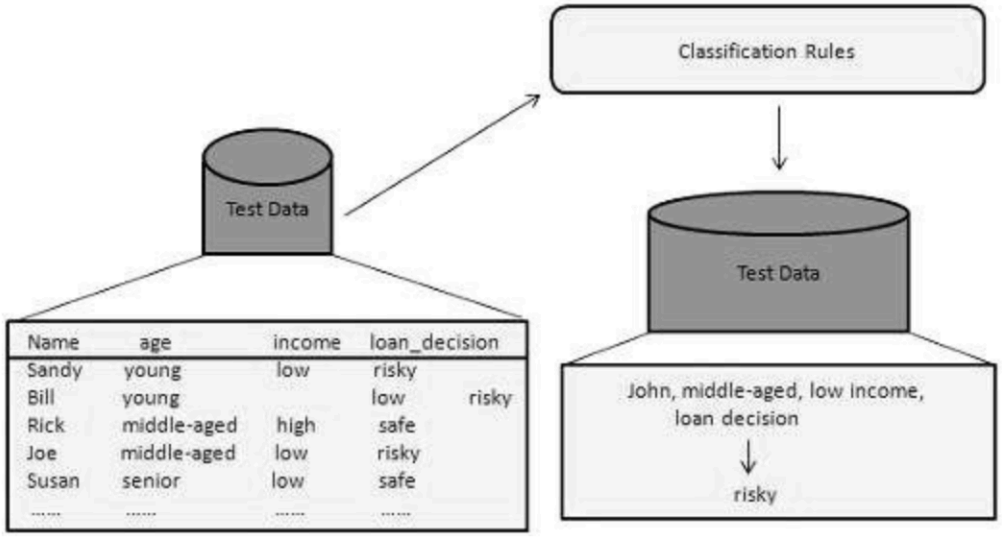
- Um conjunto de técnicas para identificar as categorias nas quais os novos pontos de dados pertencem, com base em um conjunto de treinamento contendo pontos de dados que já foram categorizados.
- Aplicação
  - previsão do comportamento do cliente específico do segmento (por exemplo, decisões de compra, taxa de perda, taxa de consumo) em que há uma hipótese clara ou resultado objetivo.
- Essas técnicas são frequentemente descritas como **aprendizado supervisionado** devido à existência de um conjunto de treinamento.

Classification

- Construir o modelo de classificação



- Usar o modelo de classificação para classificar



https://www.tutorialspoint.com/data_mining/dm_classification_prediction.htm

Cluster Analysis

- Um método estatístico para classificar objetos que dividem um grupo diverso em grupos menores de objetos similares, cujas características de similaridade não são conhecidas de antemão.
- Aplicação
  - Um exemplo de análise de cluster é segmentar consumidores em grupos auto-similares para marketing direcionado.
- Este é um tipo de **aprendizado não supervisionado** porque os dados de treinamento não são usados.

Crowdsourcing

- Uma técnica para coletar dados enviados por um grande grupo de pessoas ou comunidade (ou seja, o "crowd") através de uma chamada aberta, geralmente através de meios de comunicação em rede, como a Web.
- Este é um tipo de colaboração em massa e uma instância do uso da Web 4.0 (móvel e ubíqua).
- Exemplos
  - Crowdfunding
    - Crowdbeer
  - Waze, Airbnb, Ushahidi, Vigilante
  - Projetos open source

Data Fusion and data Integration

- Um conjunto de técnicas que integram e analisam dados de várias fontes para desenvolver *insights* de maneiras mais eficientes e potencialmente mais precisas do que se fossem desenvolvidos pela análise de uma única fonte de dados.
- Em alguns domínios há diferenças entre os dois
  - Data Integration – junção dos dados sem redução do conjunto de dados
  - Data fusion – junção com redução ou troca de dados

Data Fusion and data Integration

- Um exemplo no domínio de aplicações geoespaciais
  - O conjunto de dados fundidos é diferente de um superconjunto combinado simples em que os pontos no conjunto de dados fundidos contêm atributos e metadados que podem não ter sido incluídos para esses pontos no conjunto de dados original.

Input Data Set α

Point	X	Y	A1	A2
α1	10	10	M	N
α2	10	30	M	N
α3	30	10	M	N
α4	30	30	M	N

Input Data Set β

Point	X	Y	B1	B2
β1	20	20	Q	R
β2	20	40	Q	R
β3	40	20	Q	R
β4	40	40	Q	R

Fused Data Set δ

Point	X	Y	A1	A2	B1	B2
δ1	10	10	M	N	*Q?*	*R?*
δ2	10	30	M	N	*Q?*	*R?*
δ3	30	10	M	N	*Q?*	*R?*
δ4	30	30	M	N	*Q?*	*R?*
δ5	20	20	*M?*	*N?*	Q	R
δ6	20	40	*M?*	*N?*	Q	R
δ7	40	20	*M?*	*N?*	Q	R
δ8	40	40	*M?*	*N?*	Q	R

https://en.wikipedia.org/wiki/Data_fusion

Data Mining

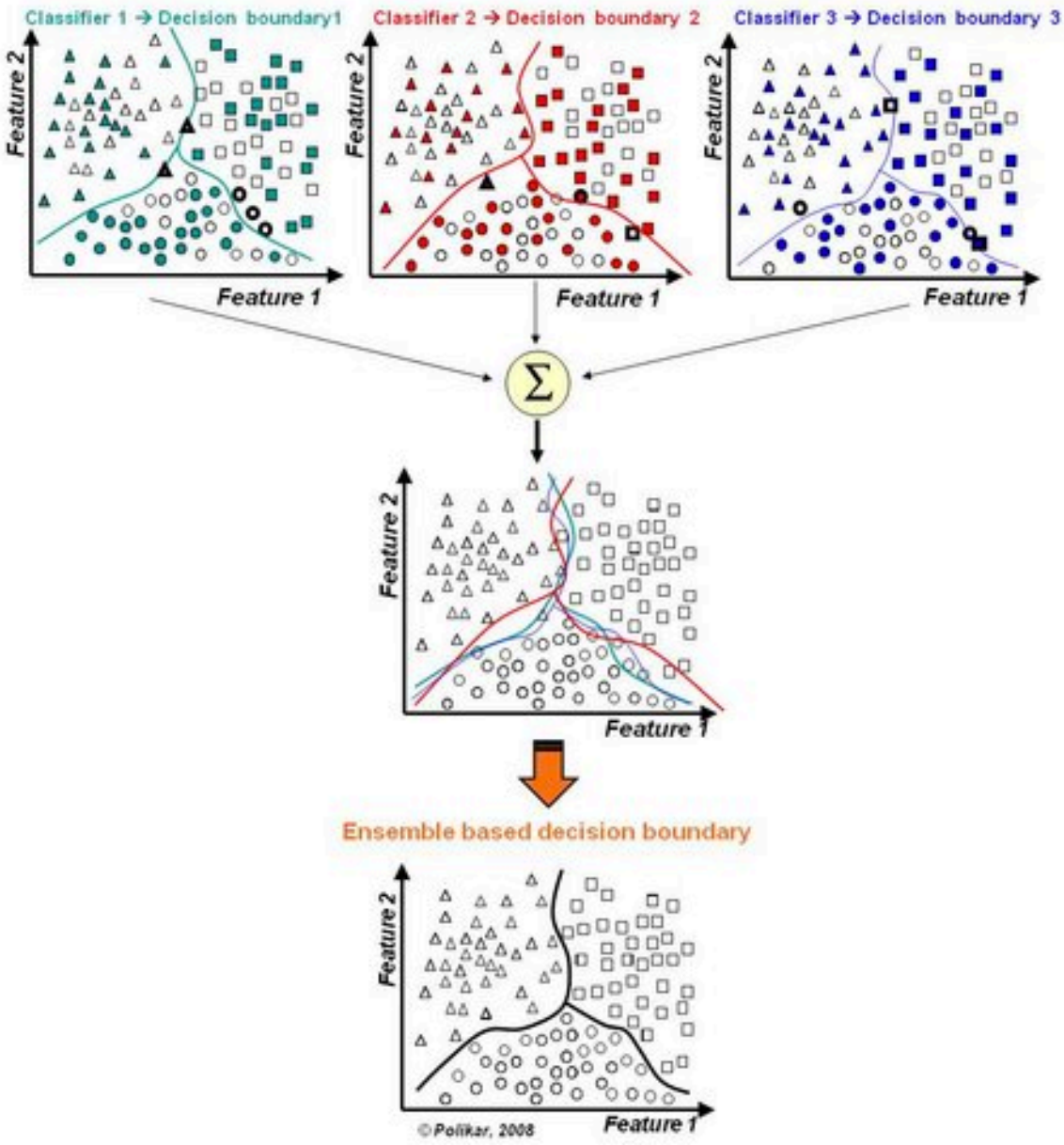
- Um conjunto de técnicas para extrair padrões de grandes conjuntos de dados, combinando **métodos de estatística** e **aprendizado de máquina** com gerenciamento de banco de dados.
- Essas técnicas incluem o **aprendizado de regras de associação**, **análise de cluster**, **classificação** e **regressão**.
- Aplicações
  - Minerar dados de clientes para determinar os segmentos com maior probabilidade de responder a uma oferta;
  - Minerar humana dados de recursos para identificar as características dos funcionários mais bem-sucedidos;
  - Análise de cesta de compras para modelar o comportamento de compra dos clientes.

Ensemble Learning

- Uso de múltiplos **modelos preditivos** (cada um desenvolvido usando **estatística** e / ou **aprendizado de máquina**) para obter melhor desempenho preditivo do que poderia ser obtido de qualquer um dos modelos individuais.
- Este é um tipo de **aprendizado supervisionado**.

Ensemble Learnin

- Combinando um conjunto de classificadores para reduzir erros de classificação e / ou seleção de modelos.



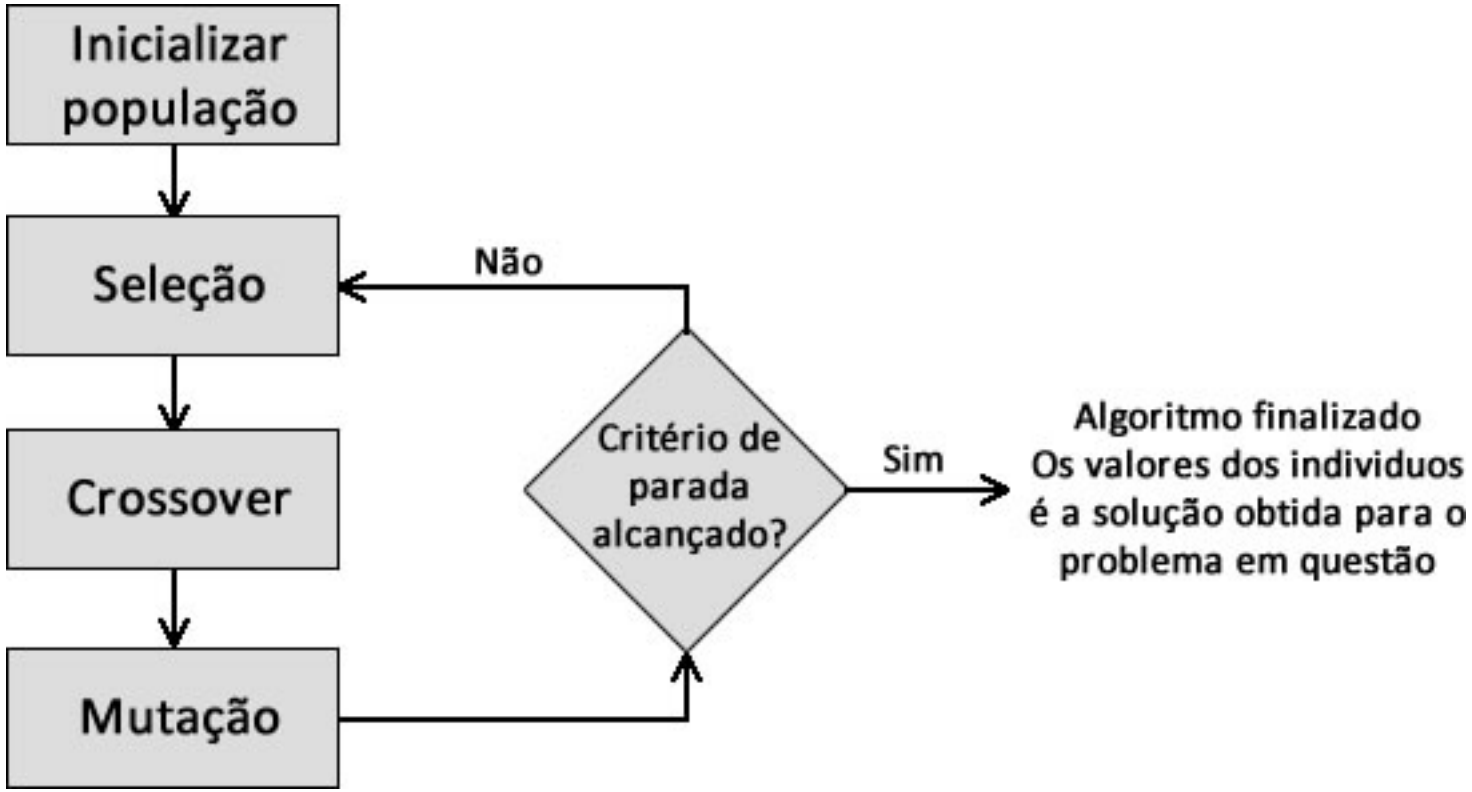
http://www.scholarpedia.org/article/Ensemble_learning

Genetic algorithms

- Uma técnica usada para **otimização** e **busca** que é inspirada no processo de evolução natural ou "sobrevivência do mais apto".
  - soluções potenciais são codificadas como "cromossomos" que podem se combinar (*crossover*) e sofrer mutações.
  - Esses cromossomos individuais são selecionados para a sobrevivência dentro de um "ambiente" modelado que determina a adequação ou desempenho de cada indivíduo na população.

Genetic algorithms

- Processo geral



http://www.computacaointeligente.com.br/algoritmos/o-algoritmo-genetico-ga/

Genetic algorithms

- Esses algoritmos são adequados para resolver problemas não-lineares.
- Exemplos de aplicações incluem:
  - melhorar o agendamento de tarefas na fabricação
  - otimizar o desempenho de uma carteira de investimentos.
  - Etc.

Optimization

- Um portfólio de técnicas numéricas usadas para redesenhar sistemas e processos complexos para melhorar seu desempenho de acordo com uma ou mais medidas objetivas (por exemplo, custo, velocidade ou confiabilidade).
- Aplicações incluem
  - melhoria de processos operacionais: agendamento, roteamento e layout do piso
  - tomar decisões estratégicas: estratégia de gama de produtos, análise de investimentos vinculados e estratégia de portfólio de P & D.
  - Algoritmos genéticos são um exemplo de uma técnica de otimização.

Machine Learning

- Uma subespecialidade inteligência artificial preocupada com o design e o desenvolvimento de algoritmos que permitem aos computadores desenvolver comportamentos baseados em dados empíricos.
- Um dos principais focos da pesquisa em aprendizado de máquina é aprender automaticamente a reconhecer padrões complexos e tomar decisões inteligentes com base em dados.

Machine Learning

- Categorias das tarefas em Machine Learning
  - Aprendizado supervisionado: Há dados de treinamento
    - Fornece Entradas e Saídas desejadas
    - Objetivo: aprender uma regra geral que mapeia as entradas para as saídas
  - Aprendizado não supervisionado: não há dados de treinamento
    - Não fornece padrões de entrada
    - Objetivo: encontrar estrutura nas entradas fornecidas. Descobrir novos padrões nos dados

Machine Learning

- Categorias das tarefas em Machine Learning
  - Aprendizado semi-supervisionado: fornece um treinamento incompleto (algumas/varias saídas desejadas incompletas)
  - Aprendizado por reforço: Um programa de computador interage com um ambiente dinâmico, em que o programa deve desempenhar determinado objetivo (por exemplo, dirigir um veículo)
    - É fornecido, feedback quanto a premiações e punições, na medida em que é navegado o espaço do problema

https://pt.wikipedia.org/wiki/Aprendizado_de_m%C3%A1quina

Pattern recognition

- Um conjunto de técnicas de aprendizado de máquina que atribui algum tipo de valor de saída (ou rótulo) a um determinado valor de entrada (ou instância) de acordo com um algoritmo específico.

Supervised learning

- O conjunto de técnicas de aprendizado de máquina que infere uma função ou relacionamento de um conjunto de dados de treinamento.

Unsupervised learning

- Um conjunto de técnicas de aprendizado de máquina que encontra estrutura oculta em dados não rotulados.
- A análise de cluster é um exemplo de aprendizado não supervisionado.

Natural Language Processing

- Um conjunto de técnicas de uma subespecialidade da ciência da computação (inteligência artificial) e linguística que usa algoritmos de computador para analisar a linguagem humana (natural).
- Muitas técnicas de NLP são tipos de aprendizado de máquina.
- Uma aplicação é usar análise de sentimentos nas mídias sociais para determinar como os possíveis clientes estão reagindo a uma propaganda.

Sentiment analysis

- Técnica que busca identificar e extrair informações subjetivas do texto de origem, que inclui
  - a identificação do recurso, aspecto ou produto sobre o qual um sentimento está sendo expresso e a determinação do tipo, "polaridade" (ou seja, positivo, negativo ou neutro) e o grau e intensidade do sentimento.
- Aplicações incluem
  - empresas usando redes sociais para verificar como clientes estão reagindo a seus produtos e ações.

Neural Networks

- Modelos computacionais, inspirados pela estrutura e funcionamento de redes neurais biológicas (ou seja, as células e conexões dentro de um cérebro), que encontra padrões nos dados
  - Envolve aprendizado supervisionado e não supervisionado
- Exemplos de aplicações incluem
  - a identificação de clientes de alto valor que correm o risco de deixar uma determinada empresa
  - identificar solicitações de seguro fraudulentas

Network Analysis

- Um conjunto de técnicas usadas para caracterizar relacionamentos entre nós discretos em um gráfico ou em uma rede.
  - Na análise de redes sociais, as conexões entre indivíduos em uma comunidade ou organização são analisadas, por exemplo, como a informação viaja, ou quem tem mais influência sobre quem
- Aplicações incluem
  - identificação de líderes de opinião-chave para direcionar o marketing
  - identificação de gargalos nos fluxos de informações corporativas

Predictive modeling

- Um conjunto de técnicas em que um modelo matemático é criado ou escolhido para melhor prever a probabilidade de um resultado.
- Aplicações incluem
  - uso de modelos preditivos para estimar a probabilidade de um cliente mudar de fornecedor
- Regressão é um exemplo das muitas técnicas de modelagem preditiva.

Regression

- Um conjunto de técnicas estatísticas para determinar como o valor da variável dependente muda quando uma ou mais variáveis independentes são modificadas.
- Geralmente usado para previsão ou previsão.
- Aplicações incluem
  - previsão de volumes de vendas com base em várias variáveis econômicas e de mercado
  - determinação de quais parâmetros de fabricação mensuráveis influenciam mais a satisfação do cliente.

Signal processing

- Um conjunto de técnicas de engenharia elétrica e matemática aplicada originalmente desenvolvidas para analisar sinais discretos e contínuos, ou seja, representações de grandezas físicas analógicas, como sinais de rádio, sons e imagens.
- Aplicações incluem
  - modelagem para análise de séries temporais ou implementação de fusão de dados para determinar uma leitura mais precisa combinando dados de um conjunto de fontes de dados menos precisas (isto é, extraindo o sinal do ruído).

Spatial analysis

- Um conjunto de técnicas que analisam as propriedades topológicas, geométricas ou geográficas codificadas em um conjunto de dados.
- Dados para análise espacial vêm de sistemas de informações geográficas (GIS)
  - Dados: endereços ou coordenadas de latitude / longitude.
- Aplicações incluem
  - Regressões espaciais: venda de produto *vs* localização

Statistics

- A ciência da coleta, organização e interpretação de dados, incluindo o design de pesquisas e experimentos.
- É frequentemente usado para fazer julgamentos sobre quais relações entre as variáveis poderiam ter
  - ocorrido por acaso (a “hipótese nula”)
  - sido resultado de algum tipo de relação causal subjacente (i.e., que são “estatisticamente significantes”) .

Simulation

- Modelando o comportamento de sistemas complexos, geralmente usados para previsão, predição e planejamento de cenários.
- Uma aplicação é avaliar a probabilidade de atingir metas financeiras, dadas as incertezas sobre o sucesso de várias iniciativas.

Time series analysis

- Conjunto de técnicas de estatísticas e processamento de sinais para analisar sequências de pontos de dados, representando valores em tempos sucessivos, para extrair características significativas dos dados.
- Aplicações incluem
  - O valor por hora de uma ação do mercado
  - o número de pacientes diagnosticados com uma determinada condição todos os dias

Visualization

- Técnicas usadas para criar imagens, diagramas ou animações para comunicar, entender e melhorar os resultados de análises de Big Data.

Quais consultas?

Consultas

- **Transacionais**: *online transaction processing* (OLTP)
  - acessam e processam parte dos dados e podem ser executadas rapidamente
- **Analíticas: *online analytical processing (OLAP)***
  - processam uma parte substancial dos dados para produzir relatórios para os analistas
- **Analíticas em tempo real: *real time analytical processing* (RTAP)**
  - processam uma parte substancial dos dados, na medida que eles estão sendo coletados, para produzir relatórios em tempo real (ou quase)

Consultas transacionais

- Consultas transacionais no Facebook[1] (peak/second)
  - pegar o perfil do usuário
  - pegar a lista de amigos
  - atualizar a lista de amigos
- Em 2010 (pico/segundo)
  - 13M consultas
  - 450M registros lidos
  - 3.5M registros modificados (< 1%)

[1]Facebook MySQL Tech Talk (http://livestre.am/rIpq) , November 2010.

Consultas analíticas

- Requerem acessar uma parte substancial dos dados
- Consultas analíticas no Facebook[1]
  - recomendar amigos
  - resumos para anunciantes
- Em 2010
  - 10,000 consultas analíticas por dia com diferentes características e requisitos

[1]Thusoo *et. al.* "Data Warehousing and Analytics Infrastructure at Facebook", SIGMOD, 2010.

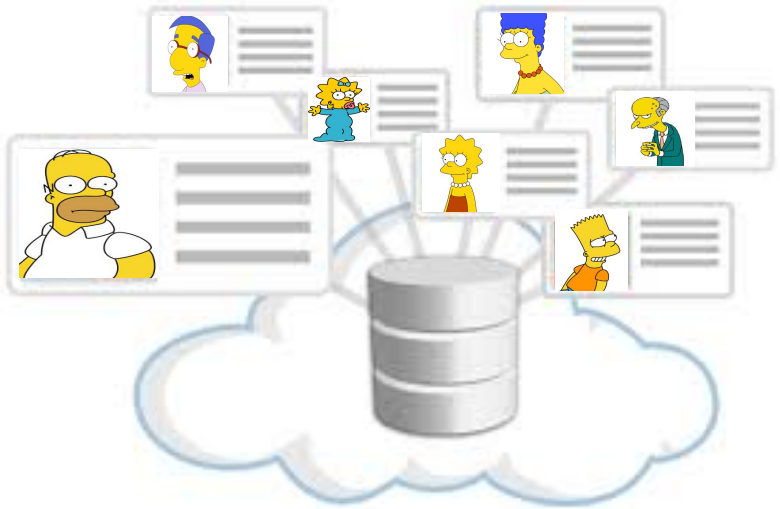
Consultas analíticas em tempo real

- Ideia
  - remover as fases de extração, transformação e carga (*ETL*)
  - fazer consultas transacionais e analíticas em um único banco
  - processar dados estruturados e não estruturados
  - requer utilizar grande quantidade de memória
- Objetivo
  - reduzir o tempo para tomada de decisão

Qual o modelo de dados?

Modelos de dados para Big Data

- *"One size does **not** fit all"*
  - são vários requisitos
  - requer a utilização de várias técnicas diferentes
  - NoSQL (*Not only SQL*)
- Modelos
  - modelo relacional
  - modelo em colunas
  - modelo chave-valor
  - modelo documentos



Modelo relacional

- Vantagens
  - ACID: Atomicidade, Consistência, Independência e Durabilidade
  - consultas em SQL
  - suporta transações
- Desvantagens
  - requer dados estruturados
  - requer esquema pré-definido
  - não suporta consulta textual de forma eficiente
  - baixa escalabilidade

Modelo em colunas

- Cada coluna é armazenada sequencialmente
- Vantagem
  - ACID
  - SQL
  - Transações
  - compactação eficiente
  - fácil integração ao modelo *MapReduce*
- Desvantagens
  - requer dados estruturados
  - não suporta busca textual de forma eficiente

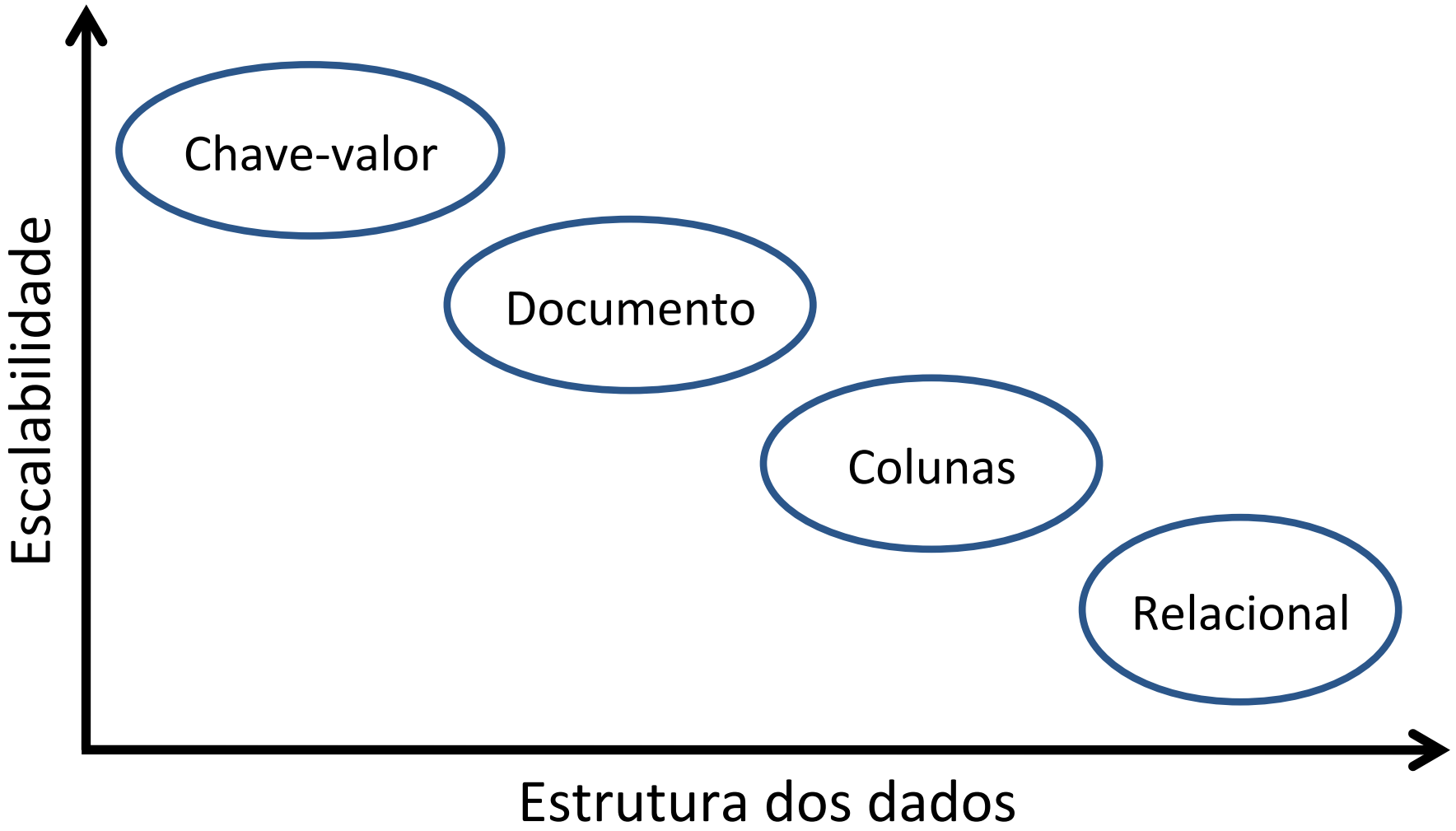
Modelo chave-valor

- Armazena pares (chave, valor)
  - similar uma Hashtable persistente
- Vantagens
  - eficiente para leitura e escrita de operações
  - alta escalabilidade
  - suporta dados não estruturados
  - busca textual
- Desvantagens
  - não suporta SQL e ACID
  - limitada a consultas por palavra chave
  - nenhuma informação sobre o valor é disponibilizada

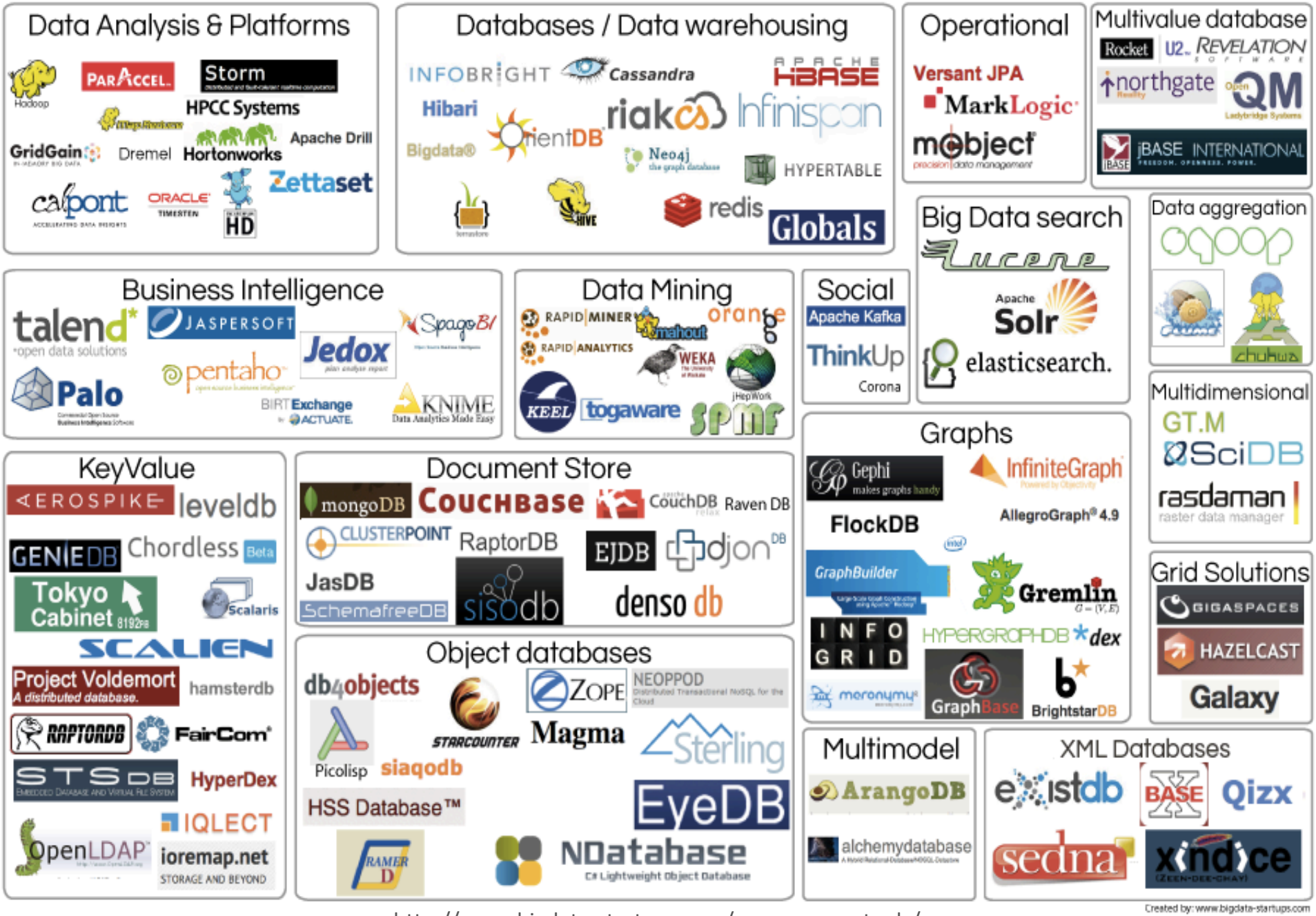
Modelo em documentos

- Similar ao modelo chave-valor
  - substitui-se valor por Documento
  - o Documento é estruturado
- Vantagens
  - mesmas vantagens da abordagem chave-valor
  - permite criar índices secundários sobre os atributos do documento
- Desvantagens
  - não suporta SQL e ACID
  - requer documento estruturado

Modelos de dados



Quais tecnologias?



http://www.bigdata-startups.com/open-source-tools/

Tecnologias para lidar com Big Data

- Há muitas tecnologias para lidar com big data, dentre elas:
  - Big Table
  - Business Intelligence
  - Cassandra
  - Cloud Computing
  - Data mart
  - Data Warehouse
  - Distributed system
  - Dynamo
  - Extract, transform, and load (ETL)
  - Google File System
  - Hadoop
  - Hbase
  - MapReduce
  - Mashup
  - Metadata
  - Non-relational databases
  - R
  - Relational Databases
  - Semi-Structured data
  - SQL
  - Stream Processing
  - Structured Data
  - Unstructured Data
  - Visualization

Tecnologias para lidar com Big Data

- **Big Table**. Sistema de banco de dados distribuído proprietário criado no sistema de arquivos do Google. Inspiração para o HBase.
- **Business Intelligence (BI)**. Um tipo de aplicação projetada para relatar, analisar e apresentar dados. As ferramentas de BI costumam ser usadas para ler dados que foram armazenados anteriormente em um **data warehouse** ou **data mart**.
  - relatórios padrão gerados periodicamente ou exibir informações sobre painéis de gerenciamento em tempo real

Tecnologias para lidar com Big Data

- **Cassandra**. Um sistema de gerenciamento de banco de dados de código aberto projetado para manipular grandes quantidades de dados em um sistema distribuído
  - originalmente desenvolvido no Facebook e agora é gerenciado como um projeto da fundação Apache Software
- **Cloud computing**. Um paradigma de computação no qual recursos de computação altamente escalonáveis, geralmente configurados como um sistema distribuído, são fornecidos como um serviço por meio de uma rede.

Tecnologias para lidar com Big Data

- **Data mart**. Subconjunto de um data warehouse, usado para fornecer dados aos usuários geralmente por meio de ferramentas de business intelligence.
- **Data Warehouse**. Banco de dados especializado otimizado para relatórios, geralmente usado para armazenar grandes quantidades de dados estruturados.
  - Os dados são carregados usando as ferramentas ETL (extrair, transformar e carregar)
  - Os relatórios geralmente são gerados usando ferramentas de business intelligence.

Tecnologias para lidar com Big Data

- **Distributed system**. Vários computadores, comunicando-se através de uma rede, costumavam resolver um problema computacional comum.
  - O problema é dividido em várias tarefas resolvidas por um ou mais computadores trabalhando em paralelo.
- Os benefícios dos sistemas distribuídos incluem
  - maior desempenho a um custo menor (cluster mais barato que supercomputador),
  - maior confiabilidade (por falta de um único ponto de falha)
  - mais escalabilidade (adicionar mais nós no cluster).

Tecnologias para lidar com Big Data

- **Dynamo**. Sistema de armazenamento de dados distribuído proprietário desenvolvido pela Amazon.
- **Extrair, transformar e carregar (ETL)**. Ferramentas de software usadas para extrair dados de fontes externas, transformá-las para atender às necessidades operacionais e carregá-las em um banco de dados ou data warehouse.
- **Google File System**. Sistema de arquivos distribuídos proprietário desenvolvido pelo Google; parte da inspiração para o Hadoop.

Tecnologias para lidar com Big Data

- **Hadoop**. Um framework software open source/livre para processamento de grandes conjuntos de dados em certos tipos de problemas em um sistema distribuído.
  - Seu desenvolvimento foi inspirado no Google MapReduce e no Google File System.
  - Foi originalmente desenvolvido no Yahoo! e agora é gerenciado como um projeto da Apache Software Foundation.
- **HBase**. Um banco de dados não-relacional de código-fonte aberto (gratuito), modelado na Big Table da Google.

Tecnologias para lidar com Big Data

- **MapReduce**. Uma estrutura de software introduzida pelo Google para processar conjuntos de dados enormes em certos tipos de problemas em um sistema distribuído.
- **Mashup**. Um aplicativo que usa e combina apresentação de dados ou funcionalidade de duas ou mais fontes para criar novos serviços.
  - Esses aplicativos geralmente são disponibilizados na Web e frequentemente usam dados acessados por meio de APIs abertas

Tecnologias para lidar com Big Data

- **Metadata**. Dados que descrevem o conteúdo e o contexto de arquivos de dados, por exemplo, meios de criação, finalidade, hora e data de criação e autor
- **Non-relational database**. Um banco de dados que não armazena dados em tabelas (linhas e colunas)
- **Relational database**. Um banco de dados constituído por uma coleção de tabelas (relações), ou seja, os dados são armazenados em linhas e colunas.
  - O SQL é a linguagem mais usada para gerenciar bancos de dados relacionais

Tecnologias para lidar com Big Data

- **R**. Uma linguagem de programação (livre) de código aberto e ambiente de software para computação estatística e gráficos.
  - A linguagem R tornou-se um padrão de fato entre os estatísticos para o desenvolvimento de software estatístico
- **Stream processing**. Tecnologias projetadas para processar grandes fluxos de dados de eventos em tempo real

Tecnologias para lidar com Big Data

- **R**. Uma linguagem de programação (livre) de código aberto e ambiente de software para computação estatística e gráficos.
  - A linguagem R tornou-se um padrão de fato entre os estatísticos para o desenvolvimento de software estatístico
- **Stream processing**. Tecnologias projetadas para processar grandes fluxos de dados de eventos em tempo real

Tecnologias para lidar com Big Data

- **Structured data**. Dados que residem em campos fixos.
  - bancos de dados relacionais ou dados em planilhas.
- **Semi-structured data**. Dados que não estão em conformidade com campos fixos, mas contêm tags e outros marcadores para separar elementos de dados.
  - texto com tags XML ou HTML.
- **Unstructured data**. Dados que não residem em campos fixos.
  - livros, artigos, corpo de mensagens de e-mail, áudio não marcado, imagem e dados de vídeo.

Tecnologias para lidar com Big Data

- **Visualization**. Técnicas usadas para criar imagens, diagramas ou animações para comunicar, entender e melhorar os resultados de análises de Big Data.


hadoop

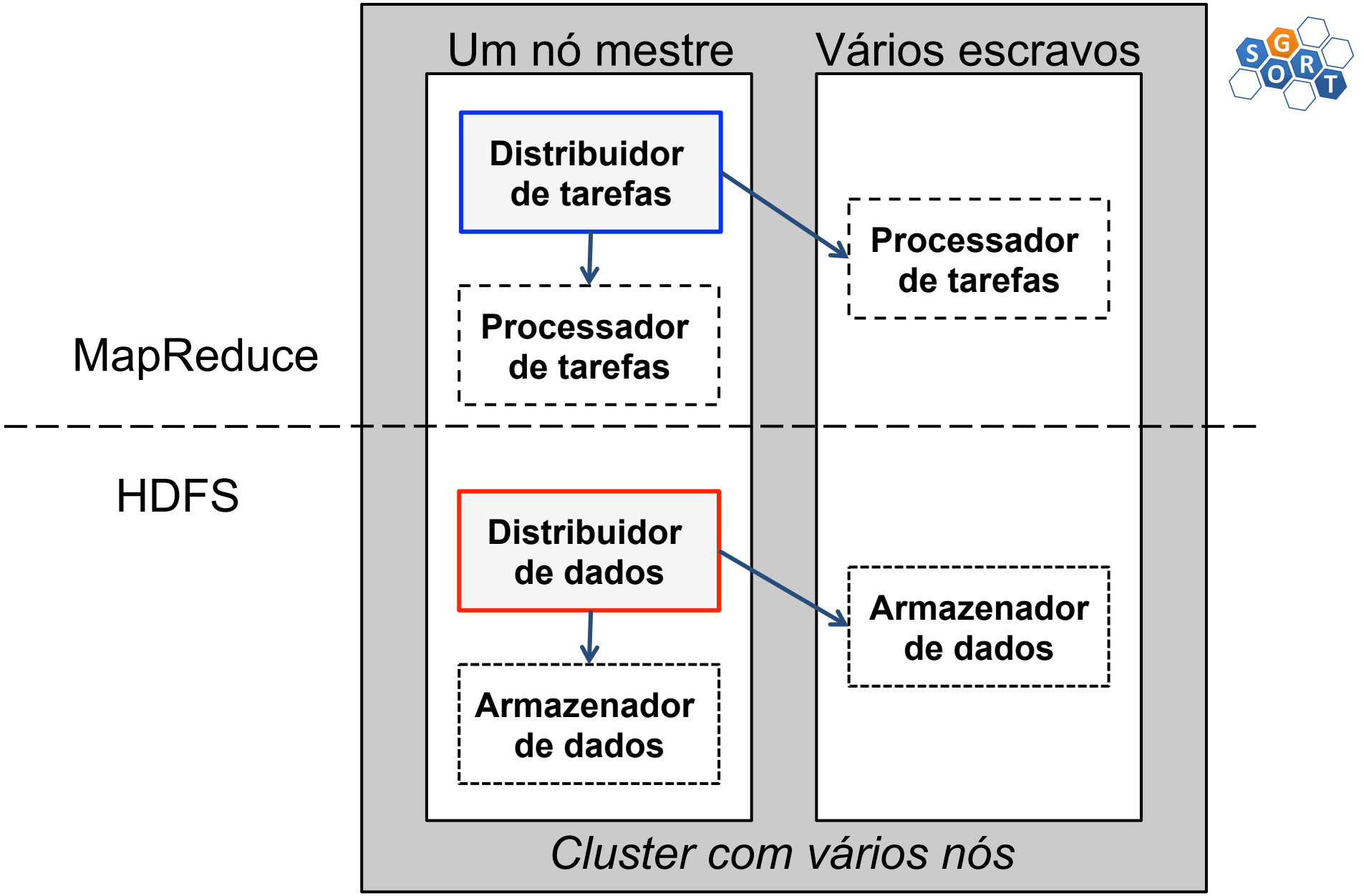
Hadoop

- Definição

  "Framework para o processamento distribuído de grandes bases de dados em um cluster de computadores, utilizando um modelo de programação funcional"

- Composto por dois componentes principais
  - sistema de arquivos distribuído (HDFS)
  - processador de dados (MapReduce)
    - **MAP**: divide tarefa
    - **REDUCE**: agrupa resultados

- Quem usa?
  - Amazon, Yahoo (cluster com 10k nós em 2008), Facebook (cluster com 21PB em 2010), CERN, JusBrasil, ...


Um nó mestre
Vários escravos
Distribuidor de tarefas
Processador de tarefas
Processador de tarefas
MapReduce
HDFS
Distribuidor de dados
Armazenador de dados
Armazenador de dados
Cluster com vários nós

Executando uma tarefa

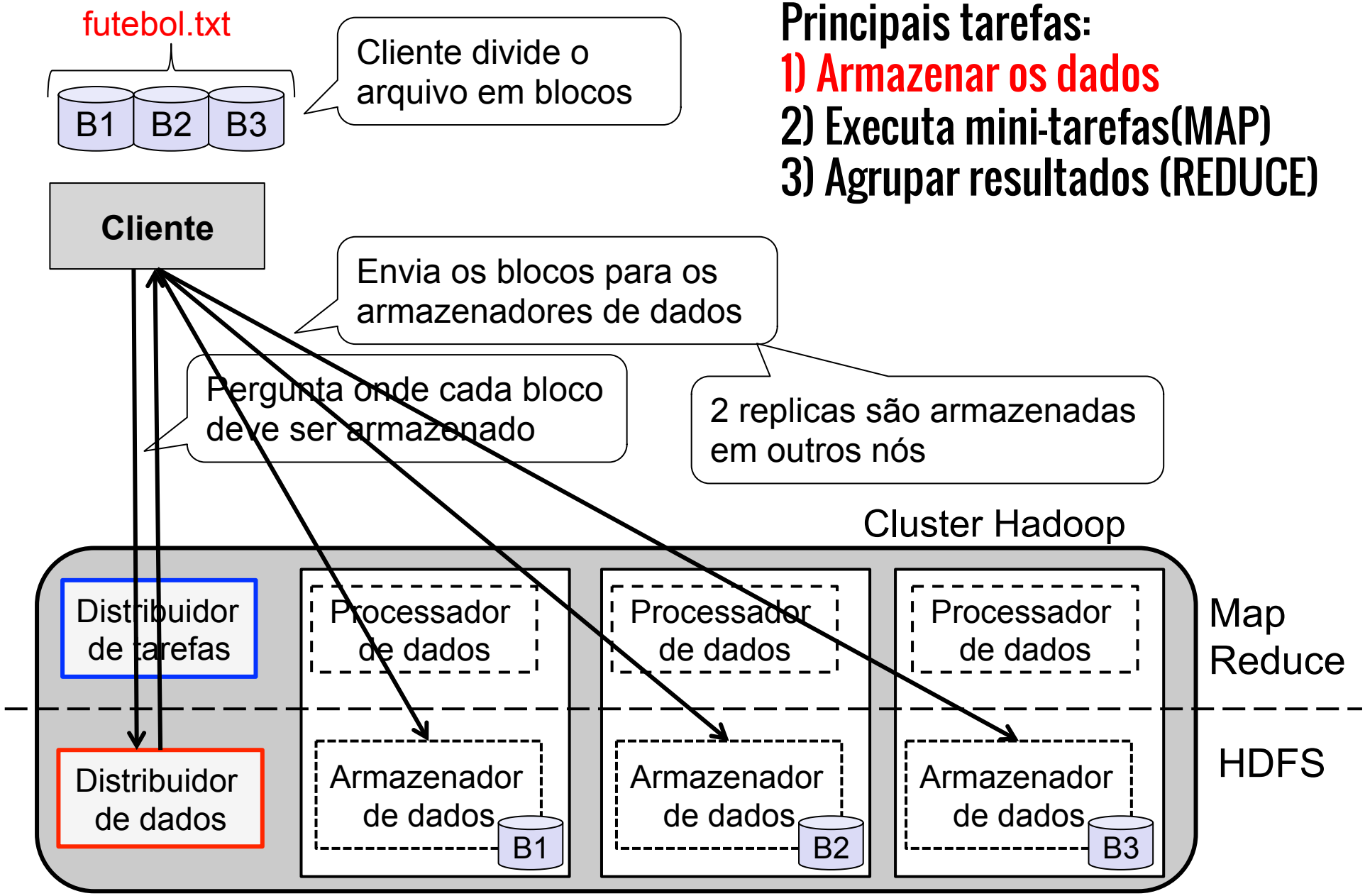
1) Armazena os dados

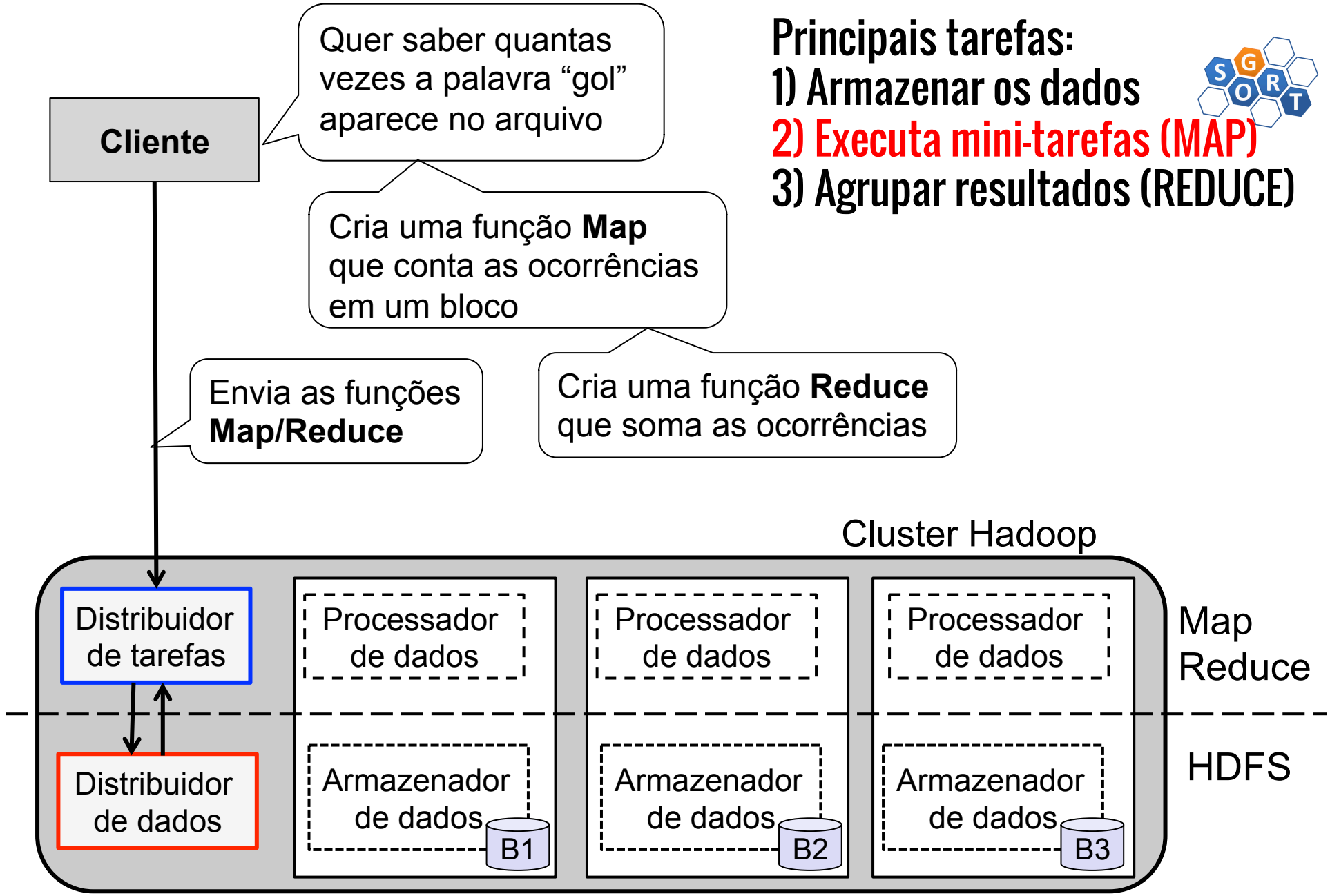
2) Executa mini-tarefas (MAP)

   - **função: Map**(K,V) → <k', v'>

3) Agrupa resultados (REDUCE)

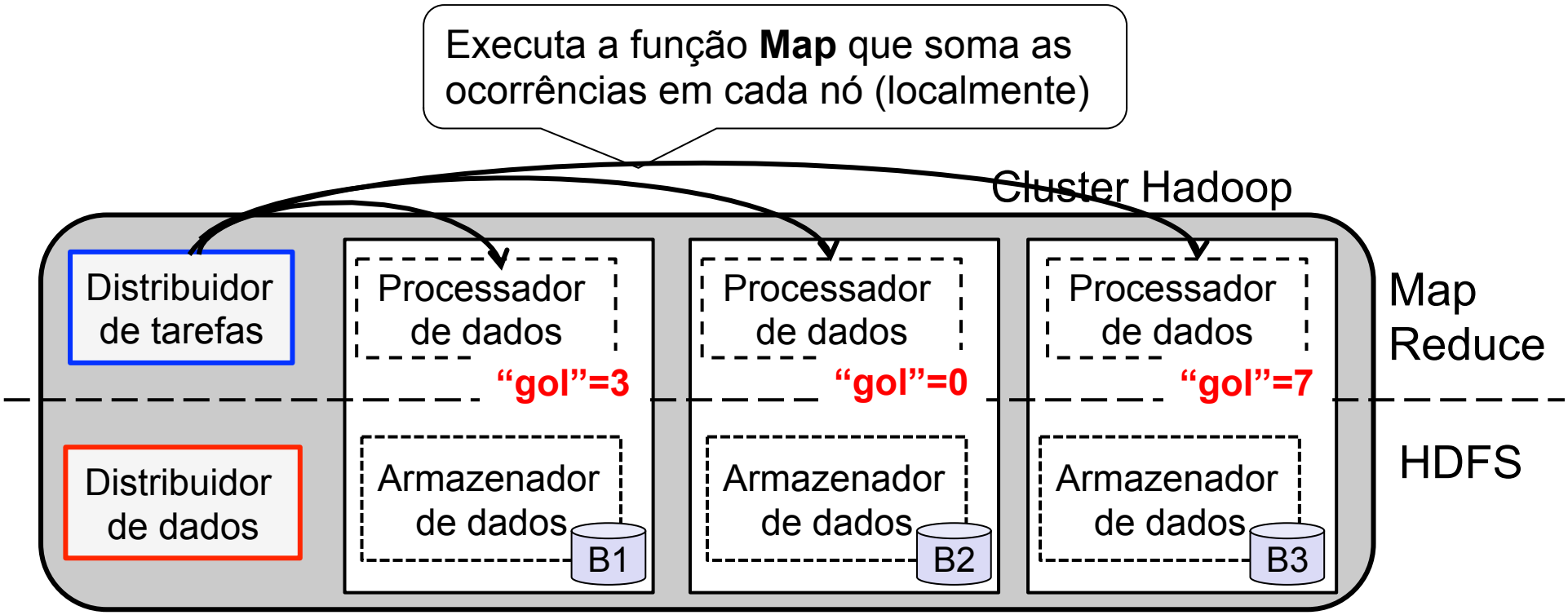
   - **função: Reduce**(k', v') → <k', v'>*


futebol.txt
B1
B2
B3
Cliente divide o arquivo em blocos
Principais tarefas:
1) Armazenar os dados
2) Executa mini-tarefas(MAP)
3) Agrupar resultados (REDUCE)
Cliente
Envia os blocos para os armazenadores de dados
Pergunta onde cada bloco deve ser armazenado
2 replicas são armazenadas em outros nós
Cluster Hadoop
Distribuidor de tarefas
Processador de dados
Processador de dados
Processador de dados
Map
Reduce
Distribuidor de dados
Armazenador de dados
B1
Armazenador de dados
B2
Armazenador de dados
B3
HDFS


Principais tarefas:
1) Armazenar os dados
2) Executa mini-tarefas (MAP)
3) Agrupar resultados (REDUCE)
Cliente
Quer saber quantas vezes a palavra "gol" aparece no arquivo
Cria uma função Map que conta as ocorrências em um bloco
Envia as funções Map/Reduce
Cria uma função Reduce que soma as ocorrências
Cluster Hadoop
Distribuidor de tarefas
Distribuidor de dados
Processador de dados
Processador de dados
Processador de dados
Armazenador de dados
Armazenador de dados
Armazenador de dados
B1
B2
B3
Map
Reduce
HDFS

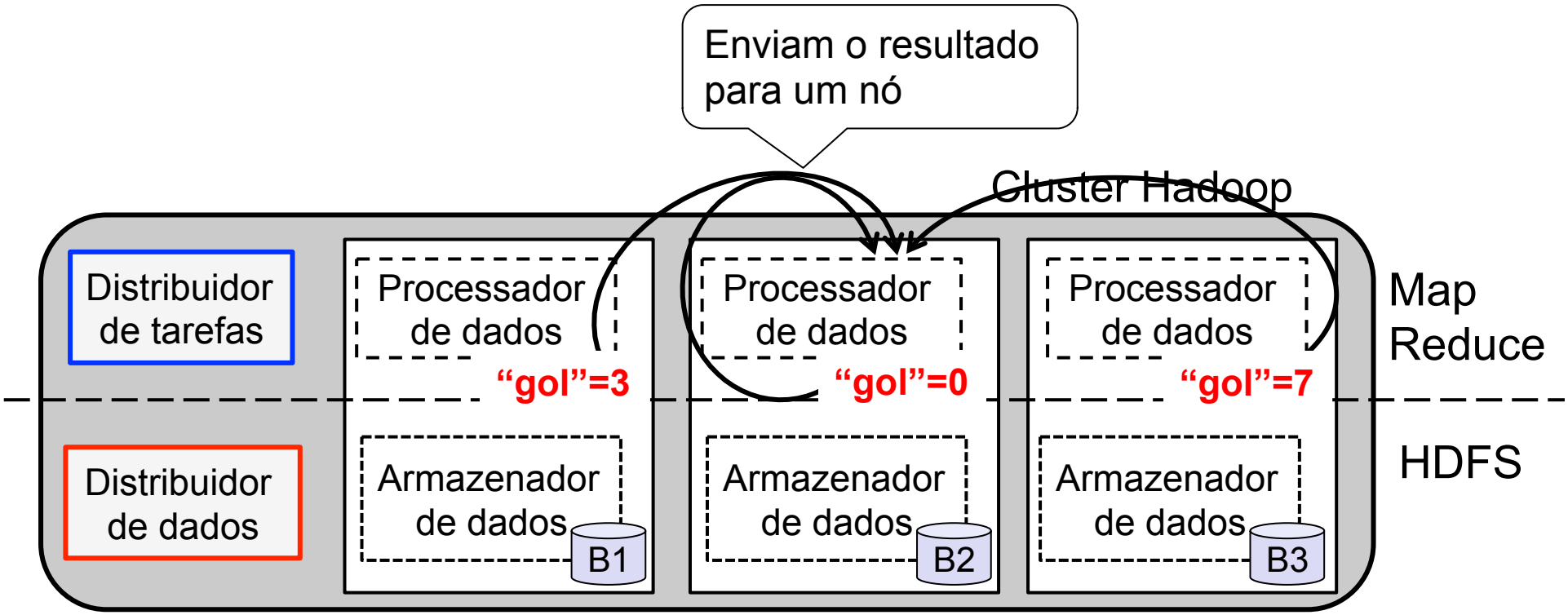
Cliente

Principais tarefas:
1) Armazenar os dados
2) Executa mini-tarefas (MAP)
3) Agrupar resultados (REDUCE)

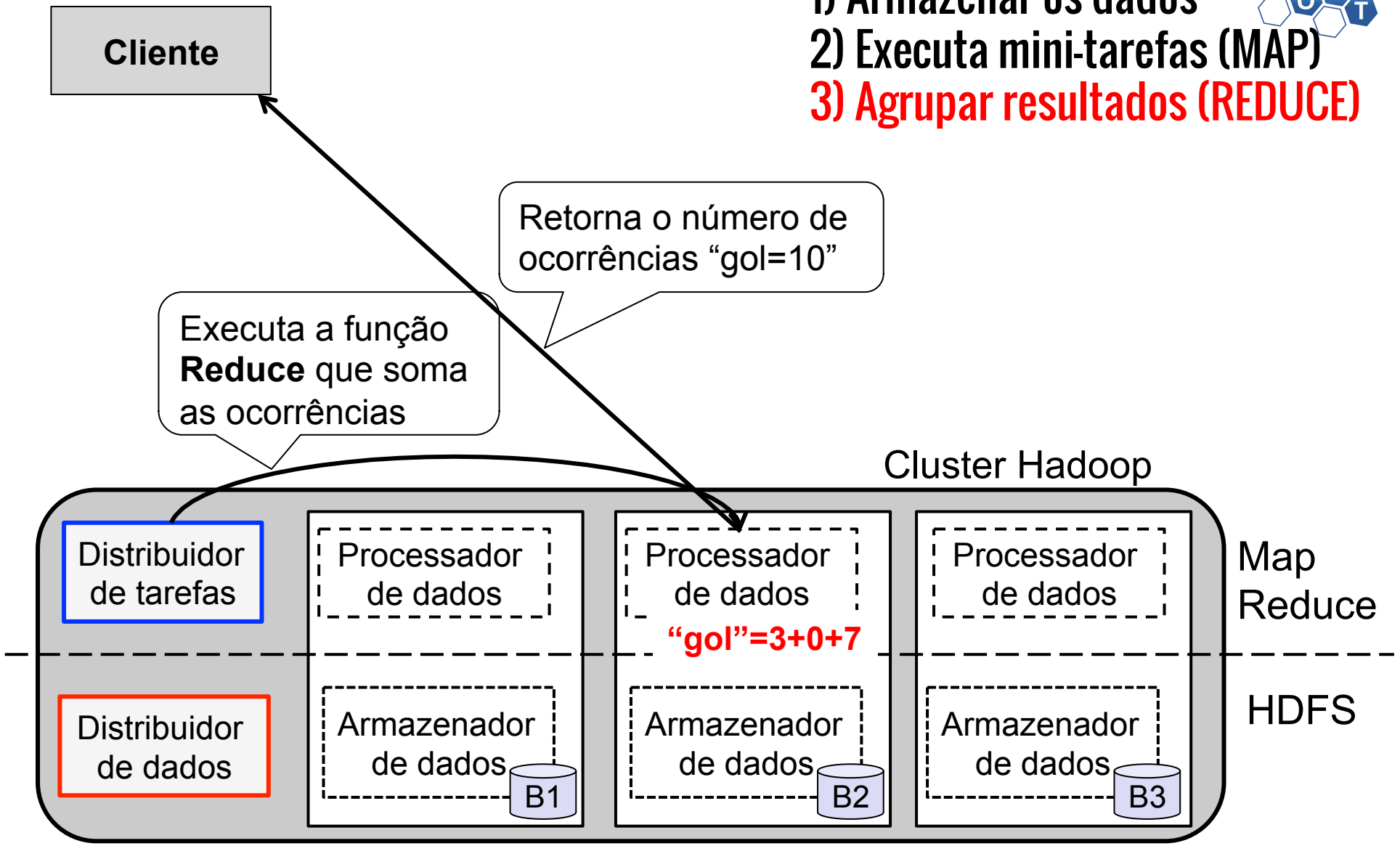


Cliente

Principais tarefas:
1) Armazenar os dados
2) Executa mini-tarefas (MAP)
3) Agrupar resultados (REDUCE)



Principais tarefas:
1) Armazenar os dados
2) Executa mini-tarefas (MAP)
3) Agrupar resultados (REDUCE)




mahout
Spark
HIVE
APACHE
HBASE
cassandra

Em resumo...

Resumo (tendências)

- Mais dados armazenados móveis
- Mais dados gerados pela Web das coisas
- Soluções aproximadas
- Armazenar os dados em memória principal
- Consultas analíticas em tempo real
- Crescimento maior dos dados estruturados
- Novas formas de visualizar os dados/resultados

Resumo

- BigData (*Buzzword*)
  - "Precisamos de ferramentas para extrair informações rapidamente de uma enorme quantidade de dados"
- Diferença
  - propriedade dos dados (5 Vs)
  - consultas preditivas, análise estatística complexa, vários tipos de dados, grandes bases de dados e consultas em tempo real
- Modelos de dados
- Consultas: OLTP, OLAP e **RTAP**
- Ferramentas: **Hadoop**

Desafios

- Consultas analíticas
  - em dados não estruturados ou incompletos
  - em tempo real
- Responder consultas analisando apenas parte dos dados (amostra)
- Volume de dados
- Mão de obra qualificada
- Ferramentas adequadas

Obrigado!

Renato Novais

renato@ifba.edu.br

Referências

- Esta Apresentação foi preparada a partir da apresentação de Big Data do Prof. João Rocha Jr. da Universidade Estadual de Feira de Santana. Obrigado João!!!

- Ananthanarayanan *et al.* "Disk-locality in datacenter computing considered irrelevant", USENIX, 2011.

- Lakshman and Malik. "Cassandra: a decentralized structured storage system", In Operating Systems Review, 2010.

- Martin Hilbert, *et al. "*The World's Technological Capacity to Store, Communicate, and Compute Information", Science 2011.

- McKinsy Global Institute. "Big data: The next frontier for innovation, competition, and productivity", 2011.

- O'Reilly Media, Inc. "Big Data Now", 2012.

- Paul C. Zikopoulos et al. "Understanding Big Data:analytics for Enterprise Class Hadoop and Streaming Data", McGraw-Hill, 2012.

- Thusoo et al. "Data warehousing and analytics infrastructure at Facebook", SIGMOD, 2010.

- Wil M.P. van der Aalst. "Data Scientist: The Engineer of the Future", Proceedings of the I-ESA Conferences, 2014.

- Wil M.P. van der Aalst. "Process Mining in the Large: A Tutorial", eBISS, 2013.