Mineração de Dados em Biologia Molecular

Big Data

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

Tópicos

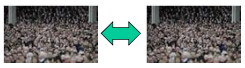
- Introdução
- Explosão de dados
- Dados nunca param de serem gerados
- Fontes dos dados
- Big Data
  - Exemplos
  - Importância
  - Características
  - Mercado e tendências

Introdução

- Sem perceber, pessoas geramos dados a todo momento
  - Aplicar para um cartão de fidelidade
    - Empresa aérea, supermercado, ...
  - Comprar com cartão de débito ou crédito
  - Navegar na internet
  - Ir ao médico
- Esses dados são armazenados em computadores (pessoais ou nuvens)

Explosão de dados

- Prática anterior
  - Poucas empresas geravam dados
  - Todo o resto (empresas e pessoas) consumia dados
- Prática atual
  - Todo mundo produz dados
  - Todo mundo consome dados



Explosão de dados

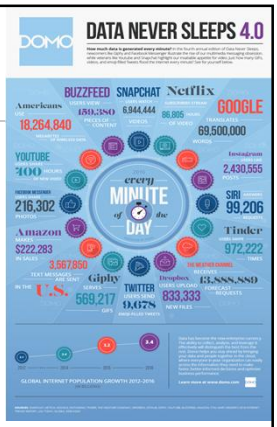
- Máquinas e pessoas continuamente geram, coletam e processam dados

Dados nunca dormem

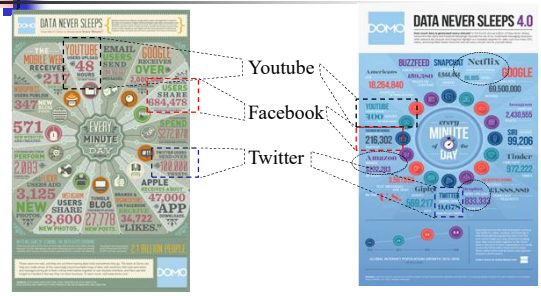
Quantos dados são gerados a cada minuto

Origem: *Domo business management platform*

https://www.domo.com



Dados nunca dormem

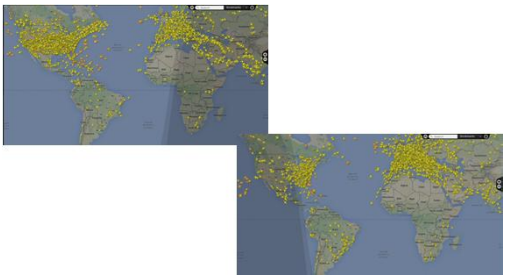


https://www.domo.com 07/2013 a 07/2016

Dados nunca dormem (manhã)

Os dados nunca dormem (tarde)

Os dados nunca dormem



Dados nunca dormem

http://www.flightradar24.com

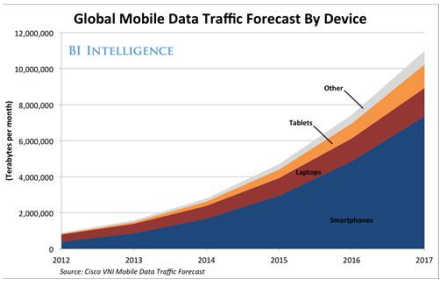
De onde vêm os dados?

- Máquinas (computadores e smartphones)
  - Logs de servidores de aplicações
    - Jogos e web sites
  - Sensores
    - Clima, água, corpo humano
  - Imagens e vídeos
    - Trânsito, câmeras de segurança

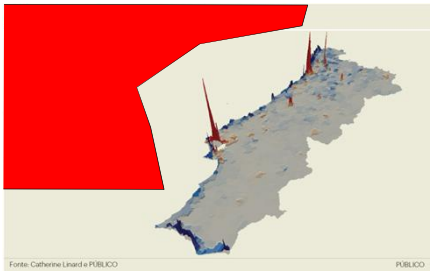
De onde vêm os dados?

- Seres humanos
  - Twitter
  - Blogs/comentários/emails
  - Compartilhamento de fotos e vídeos
  - Redes sociais
    - Facebook, linkedin, ...

De onde vêm os dados?

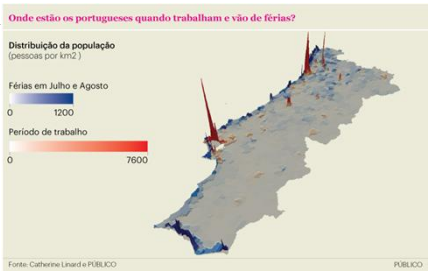


Dados gerados por ...



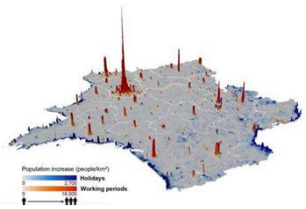
Dados de smartphones

Portugal



Dados de smartphones

França



Population dynamics between the main holiday period (July and August) and working periods in France.
Credit: Catherine Linard

http://phys.org/news/2014-10-cellphone-population-density.html#jCp

E os dados biológicos?

- Biotecnologia
  - Biologia molecular
- Meio ambiente
  - Monitoramento ambiental
  - Distribuição de espécies
- Botânica
- Paleontologia
- Saúde

E os dados biológicos?

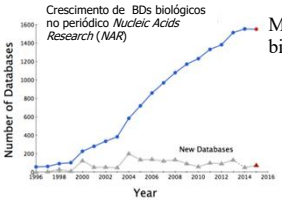
Sequencing the genome creates so much data we don't know what to do with it

Get ready for some incomprehensibly big numbers.

- Sequenciamento de dados de genoma serão a maior fonte futura de dados
  - Mais dados que astronomia, física de partículas e youtube
  - Em 2015 gerava ¼ dos dados que youtube gerava por ano

Explosão de dados biológicos

Crescimento do número de bancos de dados biológicos

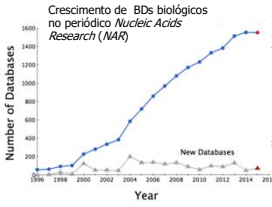


Menos Bancos de Dados (BDs) biológicos estão sendo criados?

Fonte: http://scienceblogs.com/digitalbio/2015/01/30/bio-databases-2015/

Explosão de dados biológicos

Crescimento do número de bancos de dados biológicos



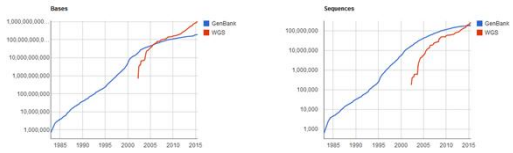
Menos Bancos de Dados (BDs) biológicos estão sendo criados?

Não

#BDs que perderam validade superou a criação de novos

Fonte: http://scienceblogs.com/digitalbio/2015/01/30/bio-databases-2015/

Explosão de dados biológicos

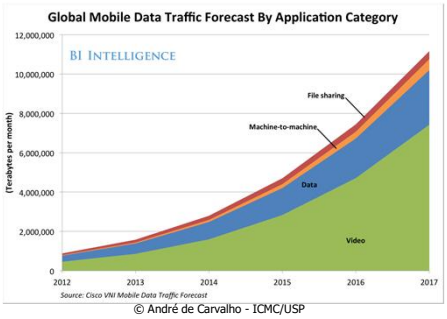


Diferentes bases de dados de sequências de nucleotídeos

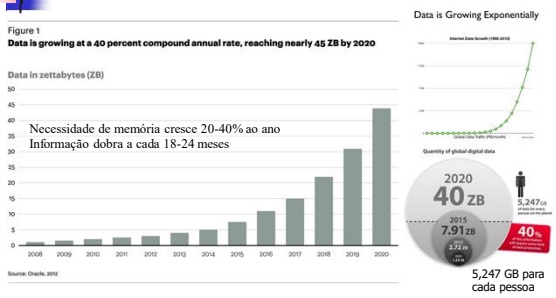
Fonte: http://www.ncbi.nlm.nih.gov/genbank/statistics

Whole Genome Shotgun (WGS): montagens de cromossomos ou genomas incompletos sequenciados pela estratégia shotgun

Que dados são esses



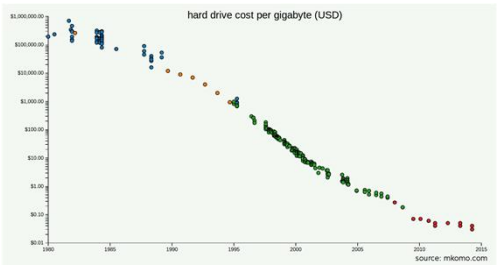
Como cresce a quantidade de dados?



Necessidade de memória cresce 20-40% ao ano
Informação dobra a cada 18-24 meses

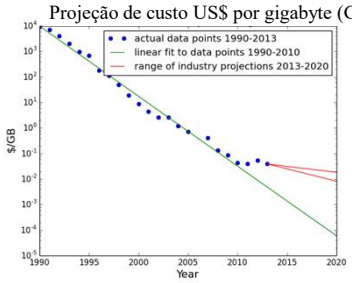
5,247 GB para cada pessoa

E o custo para armazená-los?

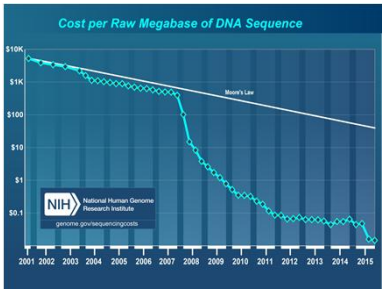


E o custo para armazená-los?

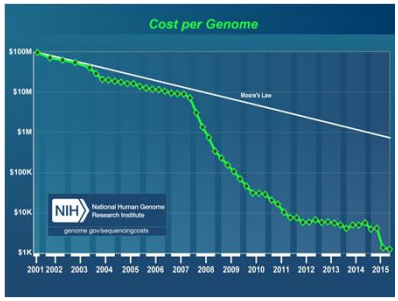
Projeção de custo US$ por gigabyte (GB)



E o custo para armazená-los?



E o custo para armazená-los?



Big Data

- Avanços recentes nas tecnologias para **aquisição**, **armazenamento** e **transmissão** de dados

cada vez mais Dados

Big Data

O que é Big Data?

O que é Big Data?

- Conjuntos de dados que são grandes demais para sistemas tradicionais de processamento de dados
- Causado por e demanda melhora contínua para:
  - Armazenamento
  - Processamento
  - Transmissão

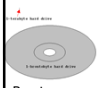
Tratamento de Big Data

- Igual a *small* data, apenas maior
  - Mas quantidade maior de dados requer abordagens diferentes
    - Arquiteturas
    - Técnicas
  - Mais dados são necessários para resolver
    - Novos problemas
    - Problemas antigos, de forma mais eficiente
  - Passo importante: KDD

Tamanho Conjunto de Dados

- Tamanhos de conjuntos de dados
  - Pequeno
    - Conjunto de dados pode ser gerenciado por ferramenta de KDD sozinha, geralmente em um único computador
  - Médio
    - Precisa da integração do ambiente de KDD com Sistemas Gerenciadores de BDs (SGBDs)
  - Grande (Big Data)
    - Quando o volume de dados é grande demais para ser gerenciado pelas ferramentas de um SGBD
    - Necessários sistemas sofisticados capazes de lidar com grandes volumes de dados estruturados e não estruturados

Tamanho Conjunto de Dados


Brontosaurus

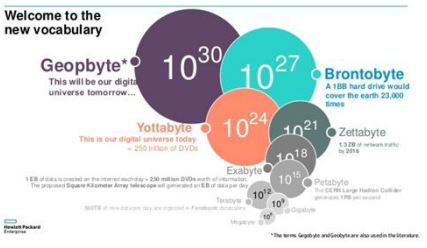
Múltiplos de Bytes				
Byte	B		10^0	
Kilobyte	KB	1024B	10^3	Pequeno
Megabyte	MB	1024KB	10^6	
Gigabyte	GB	1024MB	10^9	
Terabyte	TB	1024GB	10^{12}	
Petabyte	PB	1024TB	10^{15}	Médio
Exabyte	EB	1024PB	10^{18}	
Zettabyte	ZB	1024EB	10^{21}	Grande
Yottabyte	YB	1024ZB	10^{24}	
Brontobyte	BB	1024YB	10^{24}	
Geopbyte	GPB	1024BB	10^{24}	

E continua...

1024 Geopbytes = 1 Saganbyte
1024 Saganbytes = 1 Pijabyte
1024 Pijabytes = 1 Alphabyte
1024 Alphabytes = 1 Kryatbyte
1024 Kryatbytes = 1 Amosbyte
1024 Amosbytes = 1 Pectrolbyte
1024 Pectrolbytes = 1 Bolgerbyte
1024 Bolgerbytes = 1 Sambobyte
1024 Sambobytes = 1 Quesabyte
1024 Quesabytes = 1 Kinsabyte

1024 Kinsabytes = 1 Rutherbyte
1024 Rutherbytes = 1 Dumbnibyte
1024 Dumbnibytes = 1 Seaborgbytte
1024 Seaborgbyttes = 1 Bohrbyte
1024 Bohrbytes = 1 Hassiubyte
1024 Hassiubytes = 1 Meitnerbyte
1024 Meitnerbytes = 1 Dormstadbyte
1024 Dormstadbytes = 1 Teoentbyte

Tamanho Conjunto de Dados



Armazenamento de dados

- Computadores atuais já vêm com 1 ou 2 terabyte (TB) de memória
- Cabe em 1 petabyte (1000 TB):
  - 20 milhões de arquivos de 4 gavetas cheios
  - 500 bilhões de páginas de texto
    - Metade do conteúdo de todas as bibliotecas acadêmicas americanas combinadas
    - Papel produzido por 50 milhões de árvores
  - 7 bilhões de fotos no *facebook*

Armazenamento de dados

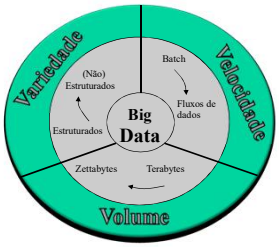
- E em 1 Terabyte (TB)?
  - 1.200 genomas humanos
  - 500 Filme de 2 horas
    - Mas até 8 dias de vídeo de segurança de alta resolução
  - 200.000 músicas em mp3
    - Dá para ouvir em ≈ 15 mil horas (quase 2 anos sem parar)

Características de Big Data

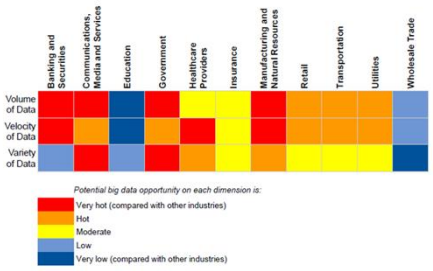
- Grande **volume** de dados, gerados a uma grande **velocidade** e com uma grande **variedade** (3 Vs)
  - Volume: tanto de dados estruturados quanto de não estruturados
  - Variedade: vindos de fontes diversas e que precisam ser integrados
  - Velocidade: gerados em fluxos cada vez mais rápidos

Características de Big Data

- Variedade:
  - Complexidade de dados
  - Dados com diferentes estruturas
    - Relacionais, Logs, textos
- Velocidade
  - Fluxos de dados em grande velocidade
- Volume
  - Escalas de Terabytes a Zetabytes (1B TBs)



Características de Big Data



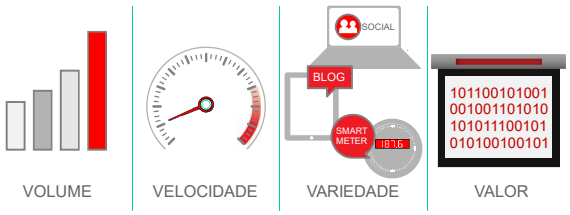
Quarto V

- Valor
  - Valor das informações contidas nos dados cresce rapidamente
    - Porém menos rapidamente que dados irrelevantes

Valor de Big Data

- Valor dos dados de 1 bilhão de perfis de usuários do facebook
  - Estimado em US$ 32 bilhões (2012) e US$ 368 bilhões (2016)
- Valor global de vendas relacionadas a aplicações de Big Data em 2012
  - Estimado em US$ 7 bilhões em 2012 e US$ 122 bilhões em 2015
    - Espera-se que cresça para mais de US$ 186 bilhões em 2019

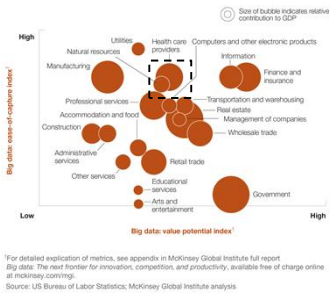
Características de Big Data



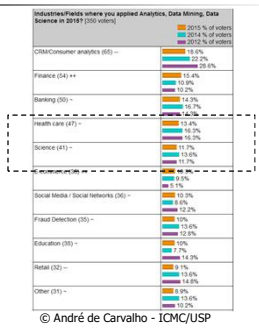
Quinto V

- Veracidade
  - 1 Em 3 tomadores de decisão não confiam nas informações que usam para decidir
    - Como usar uma informação que você não confia?
  - Um dos principais desafios de Big Data é mostrar que pode extrair informação confiável
    - Desafio aumenta com o crescimento da variedade e número de fontes

Quem se beneficiará do uso



Pesquisa Kdnuggets



Big Data não é só volume

- Quantidade de dados
- Número de fontes
- Importância
- Complexidade
- Velocidade de chegada
- Velocidade de mudança

Ambiente computacional

- Uso de mcrocomputador (notebook) para analisar dados
  - Crescimento dos três Vs dificultou análise de dados por um computador
    - Demanda maior capacidade de processamento e de armazenamento
  - Alternativas:
    - Aumentar processamento e memória do computador
      - Limite tecnológico
    - Distribuir processamento de dados entre computadores
      - Sistemas (de computação) distribuídos

Sistemas distribuídos

- Distribui uma tarefa de análise entre vários computadores conectados
  - Cluster de computadores (nós)
    - Não confundir com *clustering* (agrupamento)
- Programador precisava:
  - Sincronizar trabalho dos computadores
  - Reduzir comunicação entre computadores
- Necessidade de novas ferramentas

Cluster de computadores

Ferramentas para Big Data

- Requisitos
  - Devem ser tolerantes a falha
    - Se um ou mais computadores deixar de funcionar, processamento não para
      - Sofre uma demora proporcional à perda
    - Deve ter redundância
      - Mais de uma cópia de dados e programas
  - Devem permitir inclusão e remoção de computadores de acordo com a demanda

Ferramentas para Big Data

- MapReduce
  - Uma das primeiras ferramentas para processar big data em clusters
    - Atende requisitos anteriores
    - Facilita processamento de grandes conjuntos de dados de forma paralela e distribuída
      - Quanto mais nós, mais fácil
  - Novo paradigma de programação
    - Hadoop: implementação de MapReduce em código aberto

Ferramentas para Big Data

- Facilita escalabilidade
  - Usuário não precisa se preocupar com o aumento ou redução do número de computadores
- Duas fases
  - *Map*
    - Divisão e mapeamento de tarefas entre nós do cluster com redundância
  - *Reduce*
    - Redução das várias soluções obtidas pelos nós para uma solução final única

Hadoop

- Framework de software para aplicações que usam grandes volumes de dados
- Implementa MapReduce
  - Permite análise de vários terabytes
    - Até 25.000 TB (25 petabytes)
  - Funciona para grandes clusters
    - Pode conectar até 4500 máquinas
  - Tolerância a falhas

Outras ferramentas para Big Data

- NoSQL
  - *Not only SQL*
  - Mecanismo para armazenar e recuperar dados mais simples que bancos de dados relacionais
- Spark
  - Adapta MapRduce para fluxos de dados

Sistemas distribuídos

- Custo de aquisição e manutenção de cluster de computadores
  - Custo de mão de obra especializada gerenciamento de clusters
- *Cloud computing*
  - Computação em nuvens

Computação em nuvens

- Nova tipo de rede de computadores, em que a computação ocorre usando a internet
- Utilizada por meio de uma plataforma
  - Integra e conecta hardware e software de vários computadores usando a estrutura de internet
  - Esconde complexidade e detalhes tanto dos programadores e dos usuários
    - Comunicam-se com a nuvem utilizando uma interface gráfica simples

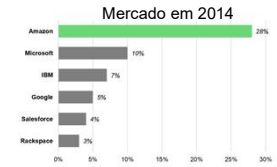
Plataforma de computação em nuvens

- Fornece serviços sob demanda
  - Sempre disponíveis, em qualquer lugar e a qualquer hora
- Usuário pode pagar de acordo com os recursos utilizados
  - Permite variar capacidades e funcionalidades
- Disponibiliza serviços para:
  - Empresas, órgãos públicos, laboratórios de pesquisa e pessoas interessadas

Computação em nuvens

Computação em nuvens

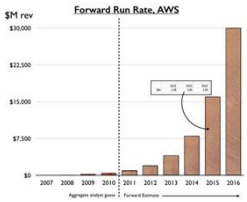
- Quatro das principais nuvens possuem ferramentas para Big Data
  - Amazon
  - Microsoft
  - Google
  - IBM



http://www.telco2research.com/articles/EB_Amazon-Web-Services-Colossal-Invincible

Expansão de nuvens

- Amazon Web Services (AWS)
  - Divisão de cloud computing da Amazon



http://blog.gardeviance.org/2015_04_01_archive.html

Nuvem da Amazon

Centro de dados

Amazon Microsoft

Expansão de nuvens

Julho de 2016

Tendências

> Até 2020, a informação será reinventada, digitalizada ou eliminada em 80% dos processos e produtos em relação a década anterior
>
> Gartner

Tendências

- Lagos de dados (*data lake*)
  - Repositório contendo grande quantidade de dados "crus" nos formatos originais
    - Permite o acesso e a análise desses dados por diferentes usuários
  - Facilita a integração das fontes (afluentes) de dados de uma organização
    - Pode capturar, misturar e explorar novos tipos de dados para extrair conhecimentos novos e de maior valor

Conclusão

- Explosão de dados
- Dados nunca param de serem gerados
- Fontes dos dados
- Big Data
  - Exemplos
  - Importância
  - Características
  - Mercado e tendências

Perguntas