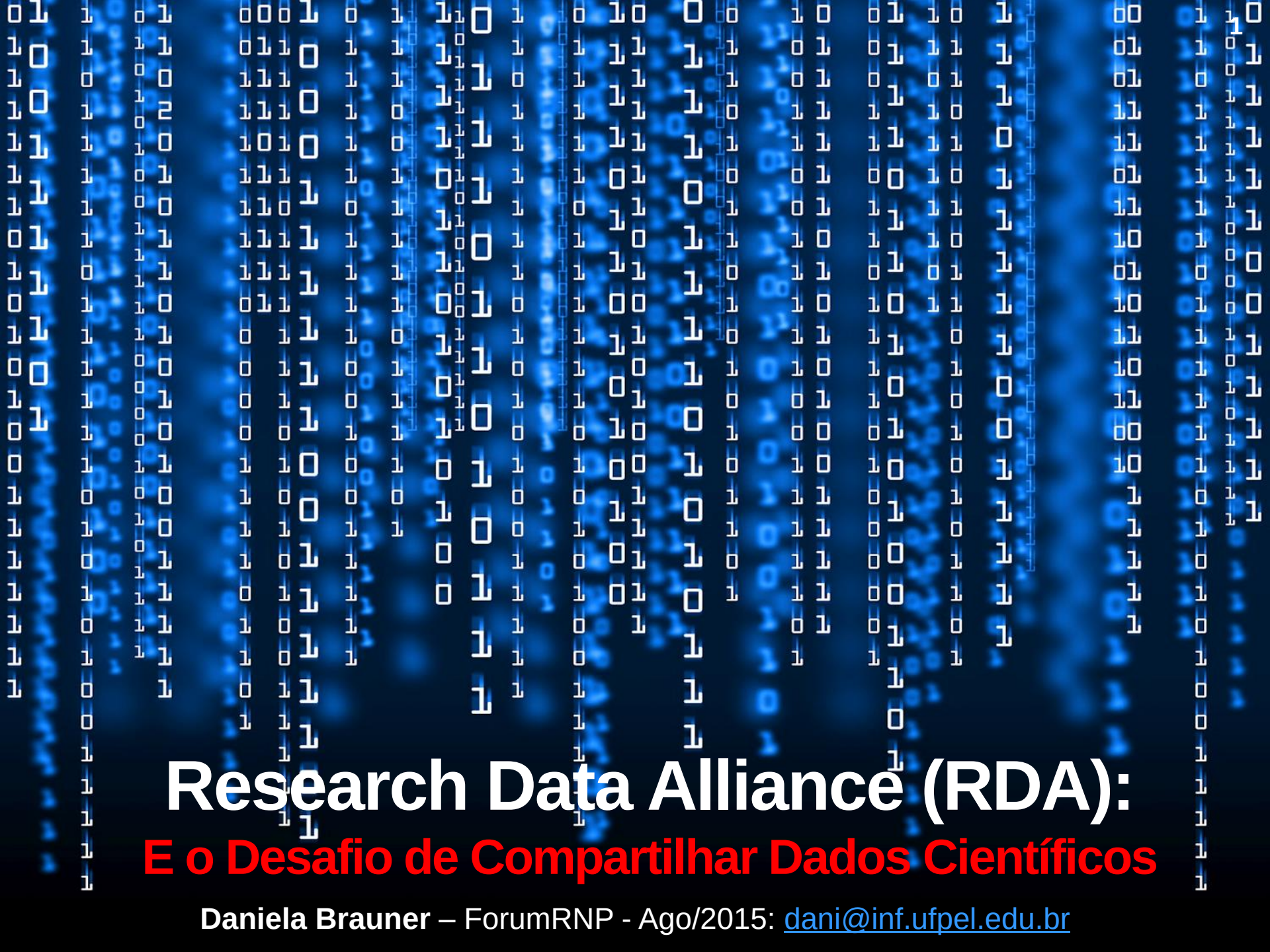
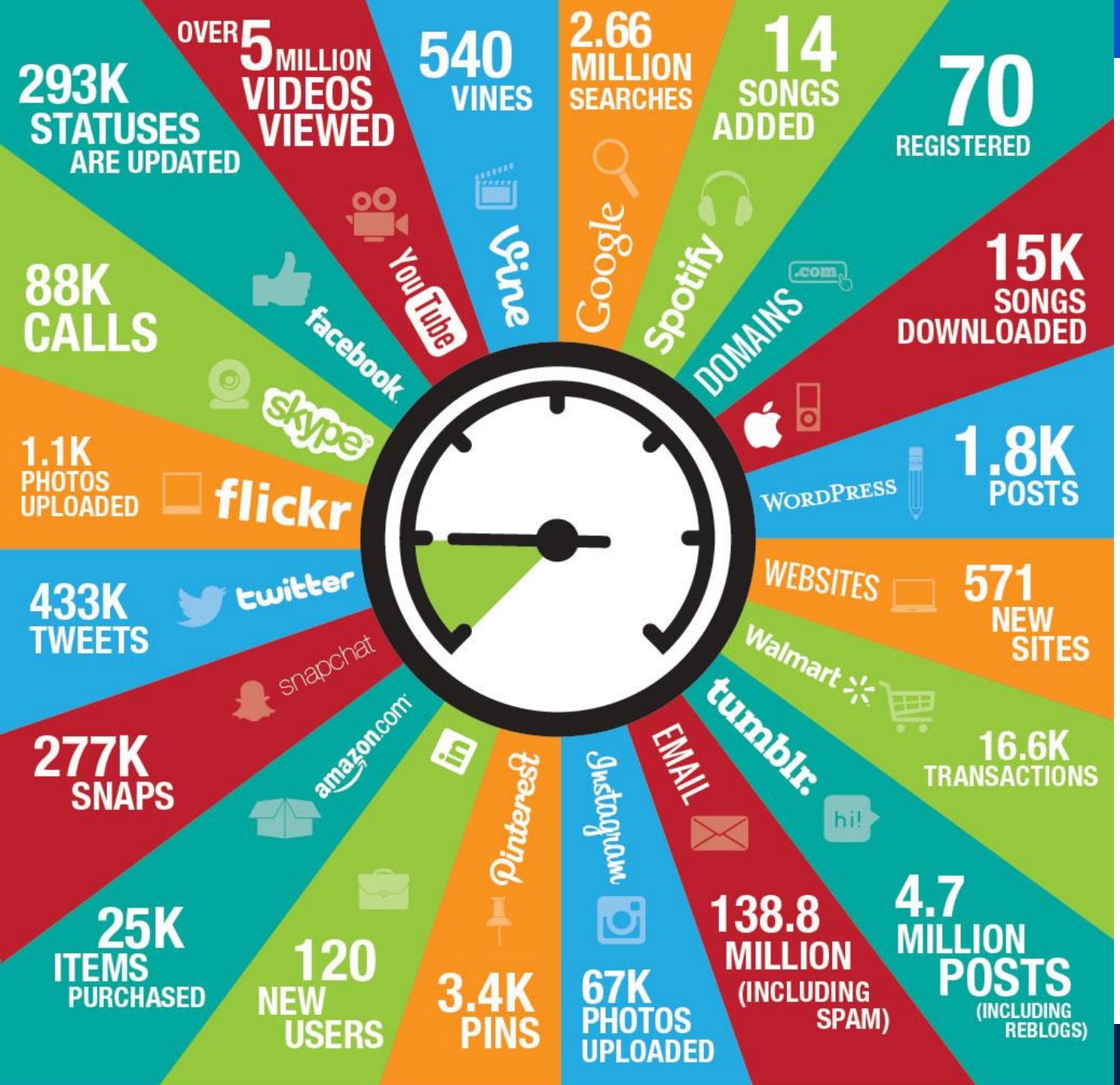


Research Data Alliance (RDA):

E o Desafio de Compartilhar Dados Científicos

Daniela Brauner – ForumRNP - Ago/2015: dani@inf.ufpel.edu.br



60 segundos ONLINE

16.6K transações
138.8 milhões de e-mails
1.1K photos adicionadas
15K músicas baixadas

…

FONTE: Qmee.
Online in 60 seconds, 2014.
http://blog.qmee.com/online-in-60-seconds-infographic-a-year-later/

Manyika, J. et al. **Big data: The next frontier for innovation, competition, and productivity**

May, 2011.

McKinsey Global Institute Report.

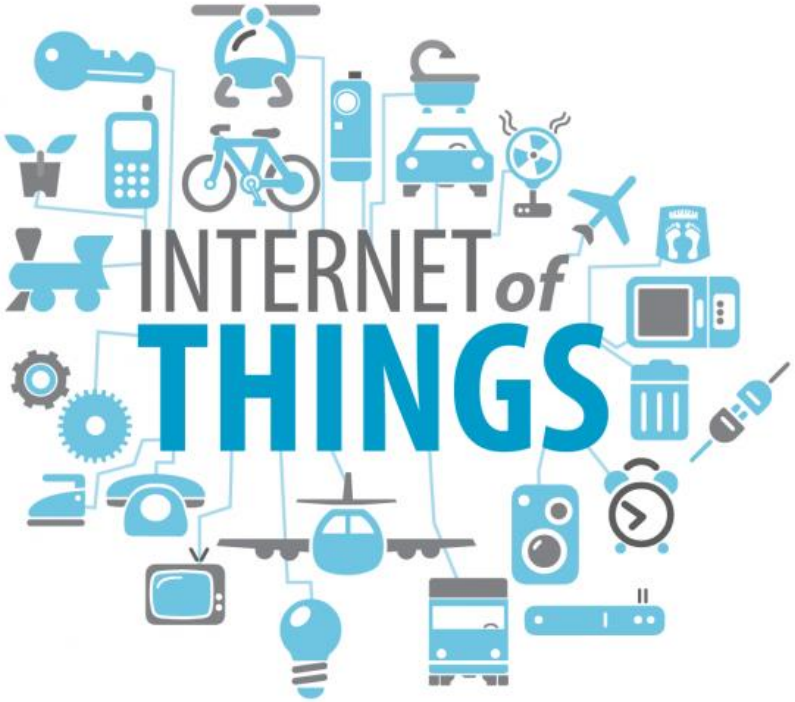
http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation

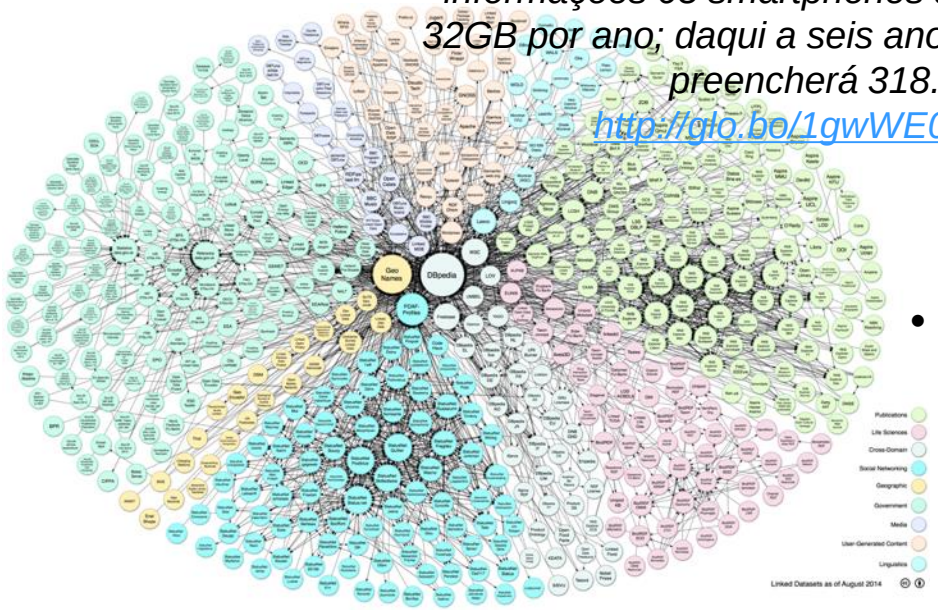
Motivação

- Milhares de sensores capturando dados por aí (*Internet of Things - IoT*).

"Carro vai gerar 25 GB de dados por hora"- http://bit.ly/1IKNT0W

"Uma família preenche com informações 65 smartphones de 32GB por ano; daqui a seis anos, preencherá 318." - http://glo.bo/1gwWE0D





- Existem milhares de coleções de dados abertos disponíveis na Web (open data)

Astronomia

- LSST - *Large Synoptic Survey Telescope*: telescópio que fará uma espécie de “filme” do céu. Coletará cerca de 2.5 milhões de visitas (filmes).
  - Estimativa em 10 anos: ~100 PB de dados coletados

Referências:
http://www.on.br
http://bravo.iag.usp.br
http://www.lsst.org

Física de Partículas

- LHC (*Large Hadron Colider*): acelerador de partículas. Coleta ~700 megabytes de dados por segundo (MB/s).
  - Estimativa em 10 anos: ~150 PB de dados coletados.

Referências:
http://home.web.cern.ch/topics/large-hadron-collider
http://opendata.cern.ch

Biodiversidade

- Grandes coleções de dados com informação sobre biodiversidade:
  - Rede *speciesLink* : ~7,7 milhões de registros com informações sobre espécimes da biodiversidade brasileira reunindo coleções de diversos herbários. Fornece ferramentas e serviços online para estimular e facilitar a publicação, acesso e uso de toda informação disponibilizada.
  - SiBBr: ~2,9 milhões de registros. Fornece uma plataforma online para estimular e facilitar a publicação, acesso e uso da informação sobre a biodiversidade brasileira.

Referências:

http://splink.cria.org.br
http://www.sibbr.gov.br

Impacto na rede: tráfego da Internet
(em volume de dados)

Ano	Tráfego da Internet Global
1992	100 GB por dia
1997	100 GB por hora
2002	100 GB por segundo (GB ps)
2007	2000 GBps
2013	28.875 GBps
2018	50.000 GBps

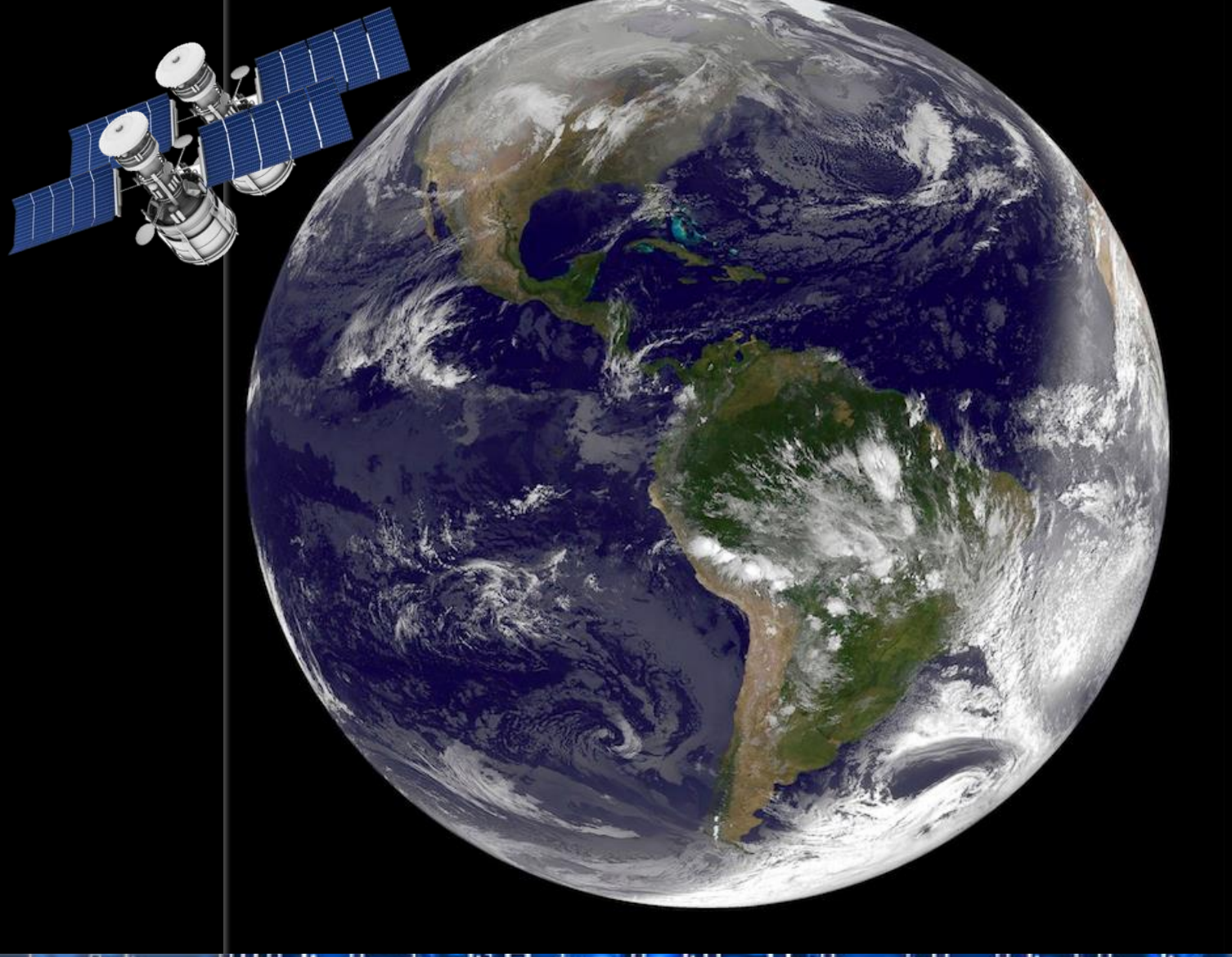
FONTE: Cisco VNI, 2014
http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/VNI_Hyperconnectivity_WP.html

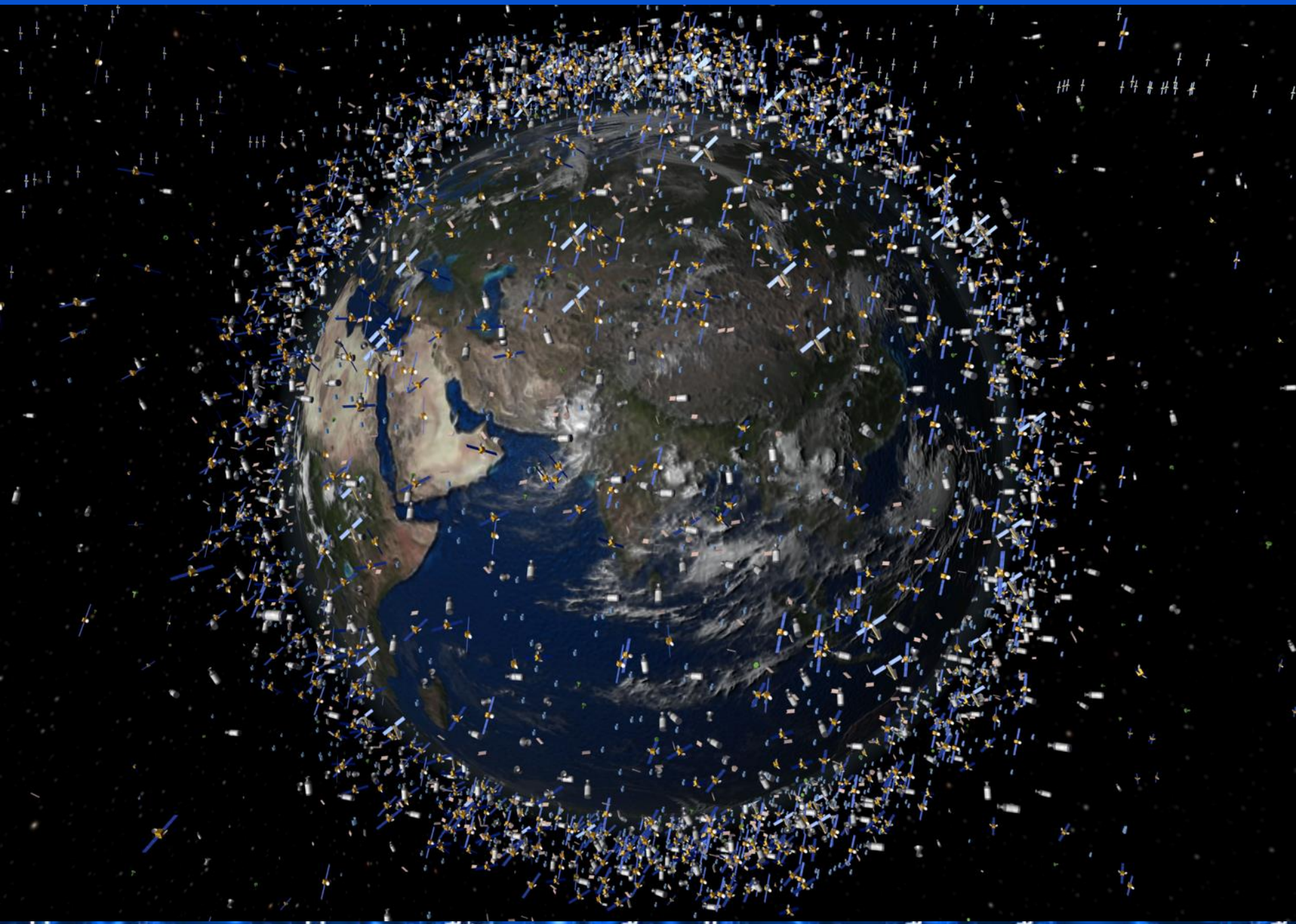
…e afinal, cadê o compartilhamento destes dados coletados no mundo?

Monitoramento Hidrossedimentológico

Sensores de pressão (nível d'água) e de turbidez.

Estação de coleta e transmissão de dados.





Motivação

- Reprodutibilidade e repetibilidade
  - Os princípios fundamentais dos métodos científicos;
    - *Reprodutibilidade:* a capacidade de reproduzir um experimento;
    - *Repetibilidade:* a capacidade de repetir um experimento;
  - A dificuldade no acesso aos dados, dificulta a reprodutibilidade da ciência (Van Noorden, 2015):
    - Tentativa de reproduzir os 50 trabalhos mais citados sobre câncer;
    - De revistas que exigem o compartilhamento de dados sob demanda;
    - Em média, a disponibilização dos dados de 31 trabalhos levou 2 meses;

FONTE: Van Noorden, R. Sluggish data sharing hampers reproducibility effort. June, 2015. In NATURE. http://www.nature.com/news/sluggish-data-sharing-hampers-reproducibility-effort-1.17694

Desafios

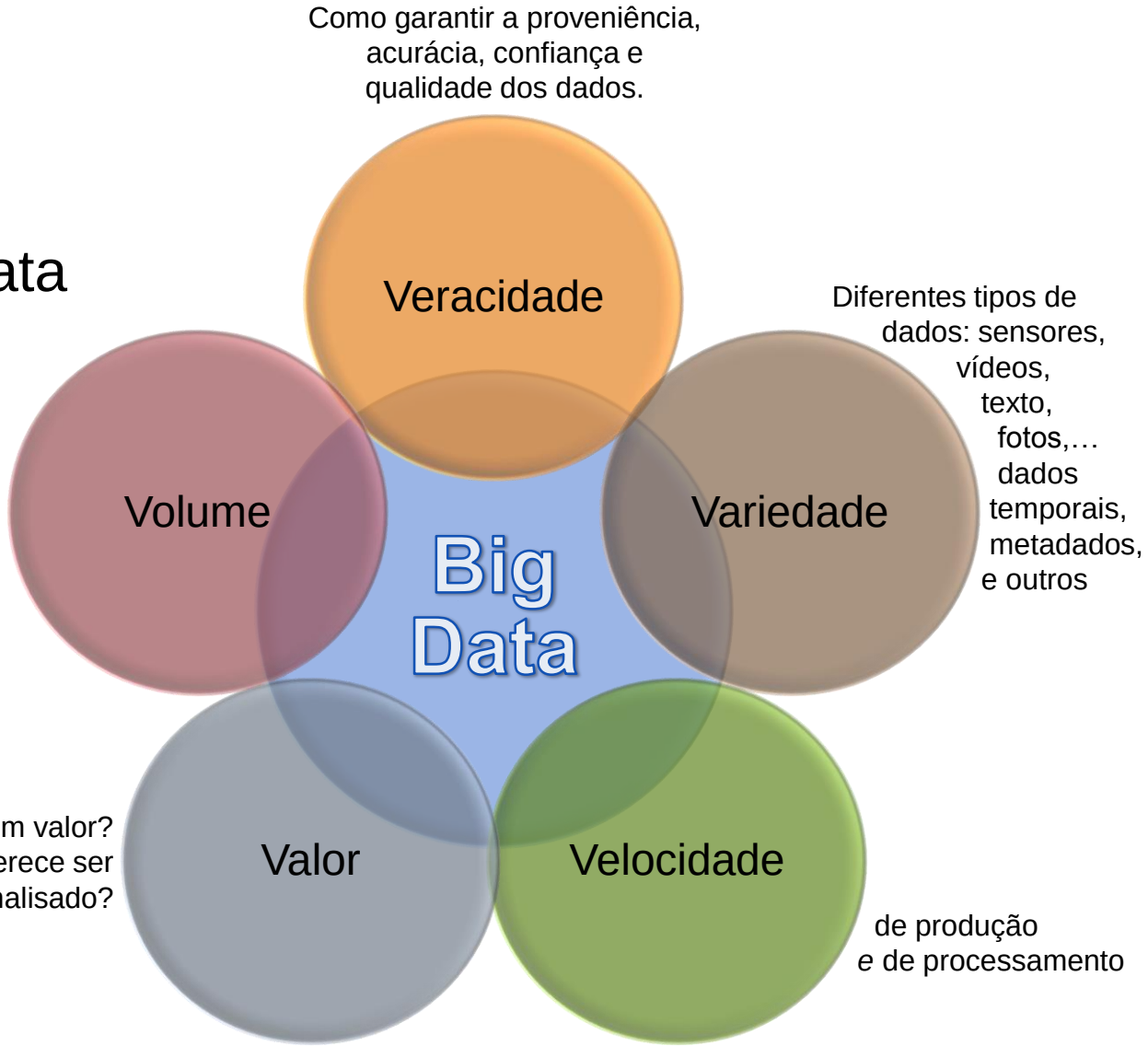
- Os acadêmicos mudam de emprego (e domicílio);
- Alguns não mantém registros de onde seus dados estão;
- Eles precisam buscar em publicações antigas como foi realizada a pesquisa;
- Eles precisam re-analisar os arquivos, que estão em formatos antigos;

Argumentos fortes indicam que os pesquisadores precisam compartilhar seus dados no momento que eles submetem um artigo (Van Noorden, 2015).

FONTE: Van Noorden, R. Sluggish data sharing hampers reproducibility effort. June, 2015. In NATURE. http://www.nature.com/news/sluggish-data-sharing-hampers-reproducibility-effort-1.17694

Desafios

- Os 5 V's da Big Data



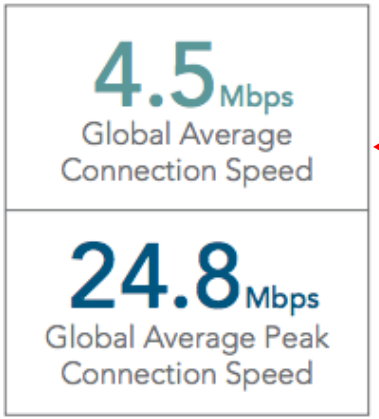
FONTE: Inspirado nos 4 V's da Big Data criado pela IBM:
http://www.ibmbigdatahub.com/sites/default/files/infographic_file/4-Vs-of-big-data.jpg

Desafios no compartilhamento de dados

- Metadados
- Identificadores persistentes
- e-Infraestrutura de compartilhamento
- e-Infraestrutura de preservação
- Certificação de repositórios
- Proveniência de dados
- Vocabulários
- Gerenciamento de dados
- Serviços de dados nacionais
- Big Data Analytics
- Disponibilidade de rede
- Garantia de citação;
- Proteção de propriedade intelectual;
- Questões de confidencialidade;
- **Sustentabilidade das e-infraestruturas**

Desafios

- Enquanto a velocidade média da Internet comercial no Brasil está abaixo da média global;



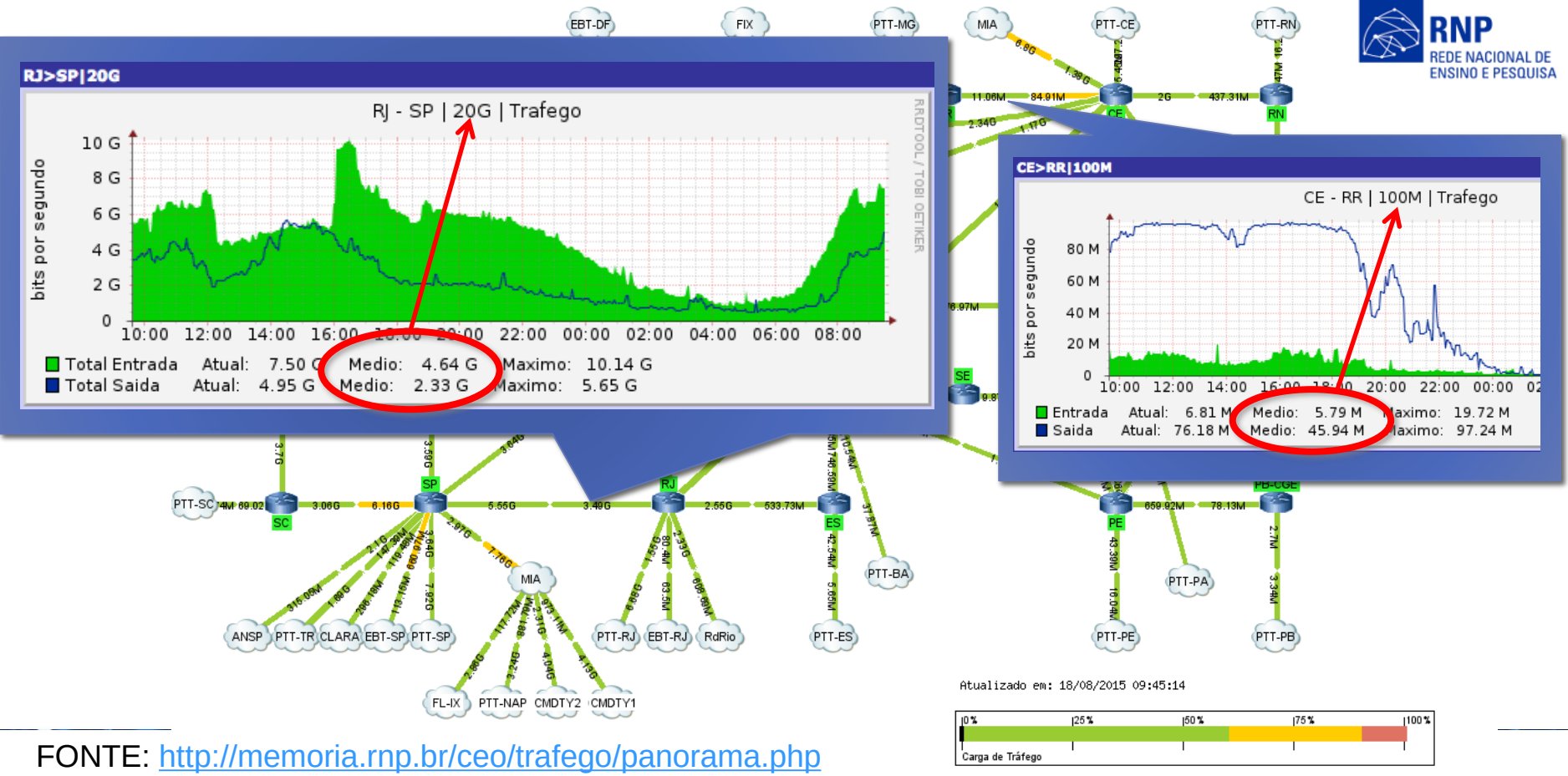
AMERICAS HIGHLIGHTS

Country/Region	Q3 '14 Avg. Mbps	Q3 '14 Peak Mbps
United States	11.5	48.8
Canada	10.3	43.7
Uruguay	5.5	58.6
Argentina	4.2	22.0
Mexico	4.1	22.8
Chile	4.1	26.1
Peru	3.6	20.6
Ecuador	3.6	20.7
Colombia	3.4	22.7
Brazil	2.9	20.5
Panama	2.9	14.2
Costa Rica	2.7	12.4
Paraguay	1.3	9.2
Venezuela	1.3	10.2
Bolivia	1.1	9.3

FONTE: Akamai State of The Internet Report- 2014. https://www.stateoftheinternet.com/downloads/pdfs/2014-q3-state-of-the-internet-report-infographic-americas.pdf

Desafios

- Podemos afirmar que temos esse mesmo problema na rede acadêmica?



FONTE: http://memoria.rnp.br/ceo/trafego/panorama.php

Ações que promovem o compartilhamento de dados

- Serviços Nacionais de Dados
- Sociedades (científicas) internacionais
  - Research Data Alliance

Ações que promovem o compartilhamento de dados

- **Serviços Nacionais de Dados**

- Sociedades (científicas) internacionais
  - Research Data Alliance

Serviços Nacionais de Dados

Oferecem serviços para **armazenamento de dados e metadados** de coleções de dados de pesquisa, para promover e facilitar a descoberta e reuso das informações. Além disso, fornecem **treinamento e assessoria em gerenciamento de dados científicos**.

- Austrália: http://ands.org.au
- Holanda: http://dans.knaw.nl
- Suécia: http://snd.gu.se/en
- Reino Unido: http://www.data-archive.ac.uk
- ...

Serviços Nacionais de Dados

- Redes de compartilhamento de dados
  - Nodos podem ser gerenciados pelas instituições participantes;
  - São necessários serviços complementares:
    - Identificadores persistentes;
    - Catálogos de metadados/vocabulários;
    - Capacitação em gerenciamento de dados;
    - Assessoria em gerenciamento de dados;

Exemplos:

- Australian Research Data Commons[1]
- Dataverse[2]

[1] http://ands.org.au/ardc.html

[2] http://dataverse.org

Ações que promovem o compartilhamento de dados

- Serviços Nacionais de Dados
- **Sociedades (científicas) internacionais**
  - **Research Data Alliance**

Research Data Alliance (RDA)
https://rd-alliance.org

- **Objetivo:** construir conexões técnicas e sociais para viabilizar o compartilhamento aberto de dados entre diferentes tecnologias, disciplinas e países, de forma a endereçar grandes desafios da sociedade em escala global.

- Criada em 2013 por um grupo de agências interessadas no tema:
  - Comissão Europeia
  - National Science Foundation
  - National Institute of Standards and Technology (NIST)
  - Australian Government's Department of Innovation

Research Data Alliance (RDA)

- Sem fins lucrativos
- Composição:

  + de 3140 indivíduos de + de 103 países

  A participação é aberta

  Get involved: https://rd-alliance.org/about/get-involved.html

Research Data Alliance (RDA)

- Governança:
  - Através de um **Conselho**:
    - Formado por indicações governamentais de diferentes países;
    - Possui atuação estratégica:
      - aconselhar os caminhos da RDA;
      - influenciar os governos locais e as agências de fomento a incluirem ações em seus planos que promovam os temas discutidos na RDA;
      - aprovar grupos de trabalhos candidatos, alinhados aos objetivos da RDA;
    - A RNP faz parte do conselho, por indicação do MCTI:
      - Michael Stanton, Diretor de Pesquisa e Desenvolvimento.

Research Data Alliance (RDA)

- **Technical Advisory Board (TAB)**: É responsável pela Estratégia Técnica do RDA, provendo o *expertise* técnico para o Conselho. Além disso auxilia no desenvolvimento e acompanhamento de Grupos de Interesse e de Trabalho:
  - Grupos de Trabalho:
    - Testar tecnologias, metodologias e elaborar recomendações
    - de curto prazo (de 12 - 18 meses)
  - Grupos de Interesse
    - Discutir e estruturar temas de interesse comum
    - de mais longo prazo

Research Data Alliance (RDA)

Temas dos grupos

- Metadados
- Identificadores persistentes
- e-Infraestrutura de preservação
- Certificação de repositórios
- Proveniência de dados
- Vocabulários
- Gerenciamento de dados
- Serviços de dados nacionais
- Big Data Analytics
- Propriedade intelectual;

…e outros.

Considerações finais

- Porque compartilhar dados?
  - Reusar dados para pesquisas;
  - Otimizar gastos com armazenamento e gestão de dados;
  - Expandir o escopo das pesquisas através de novos exploradores e oportunidades de colaborações;
  - Disseminar as pesquisas, permitindo a reprodutibilidade;
  - Preservar dados, afinal não ficam só com quem os capturou;
  - Aumentar a confiabilidade dos resultados;

Considerações finais

- Quem são os interessados em repositórios de dados e gerenciamento de dados?
  - Pesquisadores
  - Instituições
  - Indústria
  - Governo e agências de fomento
  - Editoras

Considerações finais

- O que fazer para promover o compartilhamento de dados:
  - Promover discussões sobre os desafios para o compartilhamento de dados;
  - Desenvolver a política de propriedade intelectual sobre os dados produzidos na instituição/departamento/grupo de pesquisa;
  - Requerer o depósito dos dados produzidos na instituição;
  - Requerer o depósito dos dados das publicações;
  - Requerer ações de gerenciamento de dados em chamadas de projetos;
  - Promover a adoção de ferramentas;
  - Prover e promover serviços de suporte ao compartilhamento de dados.

Obrigada!

Daniela Brauner

UFPel

dani@inf.ufpel.edu.br