WIDaT 2018

II WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

Universidade Federal da Paraíba (UFPB)

Data Science tendências e desafios

Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Grupo de Estudo, Pesquisa e Extensão em Big Data

Escola Politécnica da USP

pedro.correa@usp.br

wds.poli.usp.br

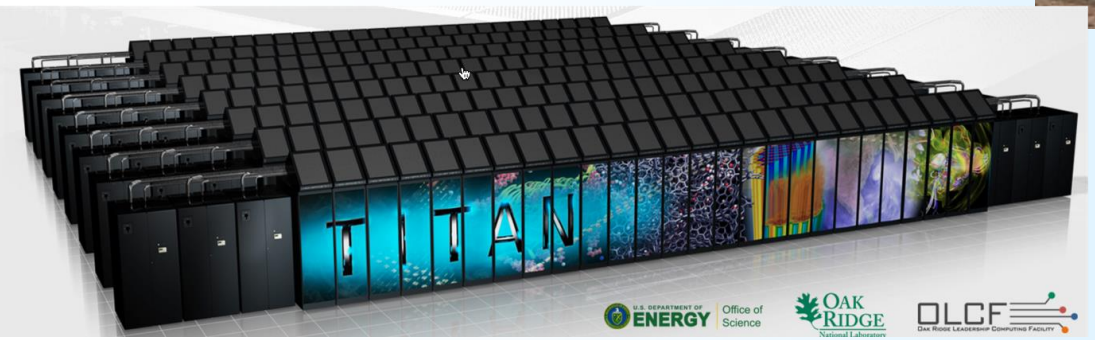
USGS science for a changing world

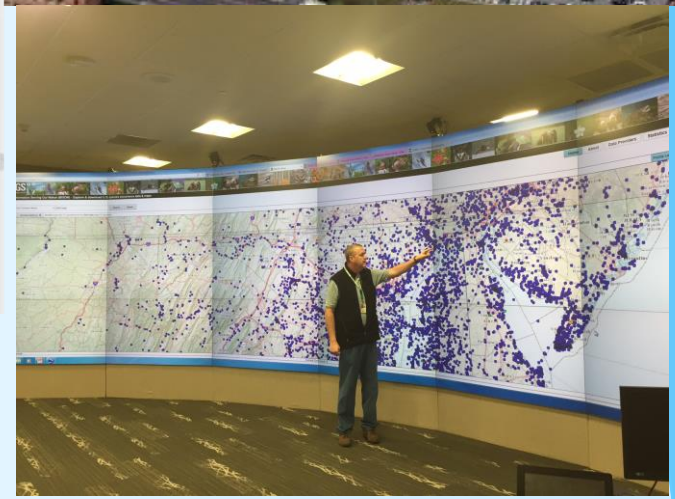
Climate Change Science Institute AT OAK RIDGE NATIONAL LABORATORY

Atuação do Grupo:

- **Projetos: FAPESP, Ministério do Meio Ambiente, ANEEL, Banco Itaú (C2D/USP);**
- **Pós-Graduação (Mestrado/Doutorado): Engenharia de Computação e *Data Science* (UT/ORNL)**
- **Cursos de Extensão:**
  - **Especialização: *Big Data* (2 anos);**
  - **Chief Data Officer (CDO)-MIT**
  - **Cursos gratuitos à comunidade.**

Colaborações Internacionais
ORNL, USGS e UT





Agenda

- Introdução;
- Gestão de Dados Científicos
- Integração/Compartilhamento de dados;
- Exemplos de projetos (ARM e DataONE);
- Conclusões;
- Referências.

1. Introdução

Conceito Dados, Informação e Conhecimento

1. **Dados:** Fluxos de fatos coletados (brutos) que representam eventos do domínio (ex: umidade, temperatura, precipitação, observação, coleta, etc);

2. **Informação:** Conjuntos de dados significativos e úteis a seres humanos em processos como o de tomada de decisões;

3. **Conhecimento:** Informações inter-relacionadas não estruturadas de regras que direcionam as tomadas de decisões.

Fonte: CORRÊA, 2011 – Adaptado Laudon, 2013

Dados – Definições

- **Dados** tem muitas definições. Mas uma que unifica nosso foco é o resultado de uma pesquisa que não está codificado nos textos de artigos, monografias, dissertações, teses, etc

- **Gestão de Dados** é uma iniciativa global que reconhece a importância dos dados primários, bem como produtos que possam ser gerados a partir dos dados.

Cortesia: Profa. Dra. Suzie Allard (University of Tennessee)

O que são dados de pesquisa?

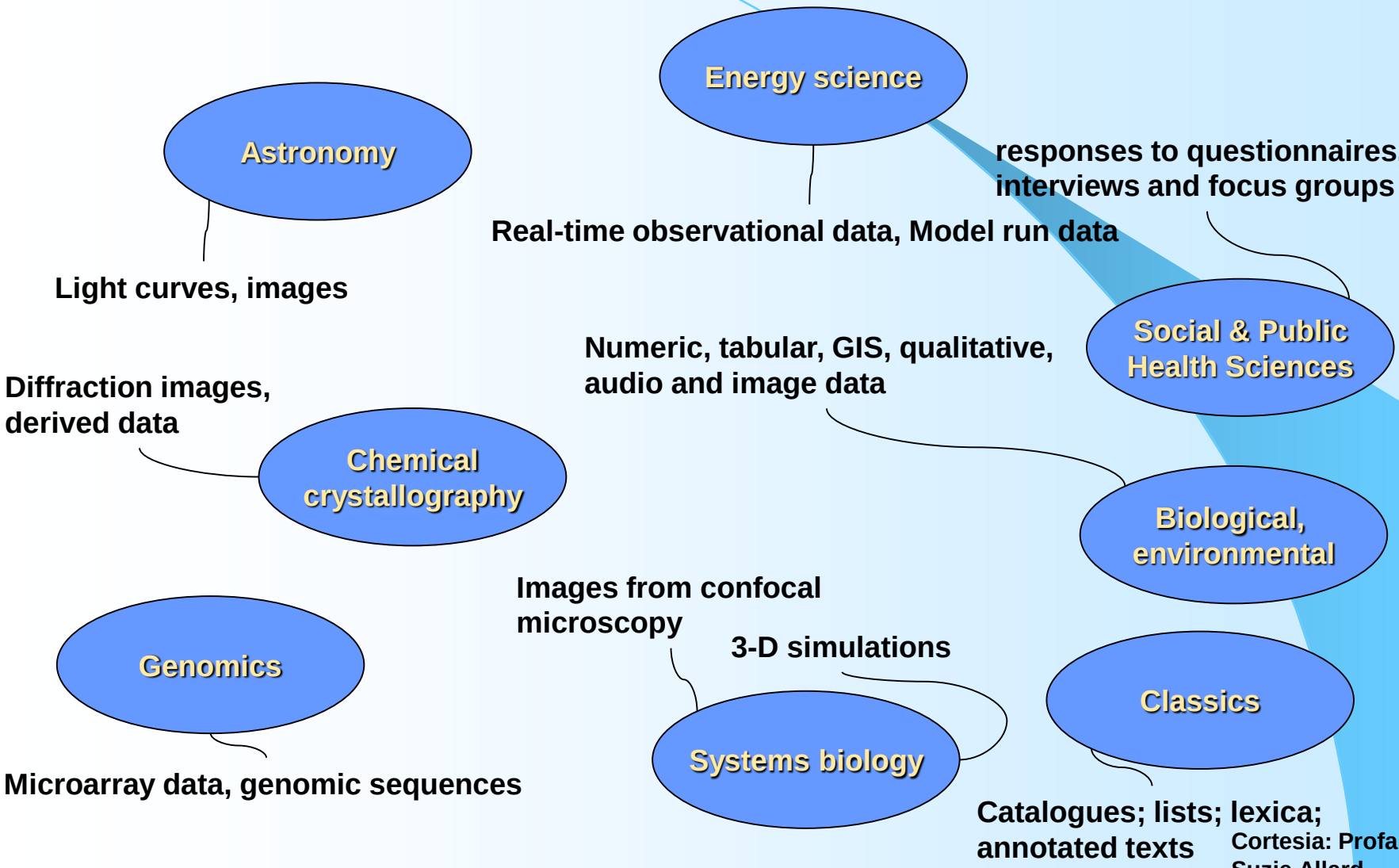
Coleções de registros ou medições utilizadas pelos pesquisadores para realizar suas pesquisas ou fornecer um registro de evidências de suas pesquisas ***"... qualquer informação que possa ser armazenada em formato digital, incluindo texto, números, imagens, vídeo ou filmes, áudio, software, algoritmos, equações, animações, modelos, simulações, etc. "***

Attributes

- **Digital**
- **Heterogeneous**
- **Contextual**
- **Valuable**

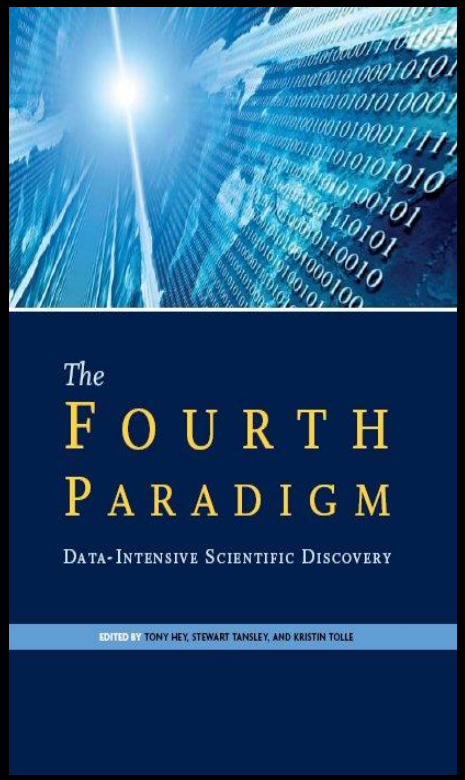
Cortesia: Profa. Dra. Suzie Allard (University of Tennessee)

Exemplos de Dados de Pesquisa



Cortesia: Profa. Dra. Suzie Allard (University of Tennessee)

Examples taken from Research Information Network (RIN) in association with the Joint Information Systems Committee (JISC) and the Natural Environment Research Council (NERC). (2008). *To Share or not to Share: Publication and Quality Assurance of Research Data Outputs Report commissioned by the Research Information Network (RIN).* Main report. June.



BIG WORLD
BIG SCIENCE
BIG DATA

Image: NASA C 98-2815

Paradigmas da ciência

Há mil anos:

→ A ciência foi **empírica**.

→ Usada para descrever fenômenos naturais.

Observações

Há poucos séculos:

→ A ciência passou a ser também **teórica**.

→ Uso de modelos, generalizações, etc.

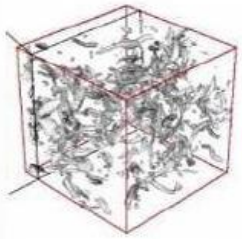
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

Leis de Kepler, Newton, Maxwell

Nas últimas décadas:

→ Pesquisadores passaram a validar seus modelos teóricos com o uso de simulações.

→ Ciência **computacional**.



Simulação de fenômenos complexos

Slides adaptados da apresentação de Jim Gray: **eScience Talk at NRC-CSTB meeting**. http://research.microsoft.com/~gray/talks

e-Science: O quarto paradigma

Hoje:

Ciência orientada a grande volume de dados (*Data-intensive Science*: Unifica teoria, experimentação e simulação).

Três atividades consideradas na exploração de dados:
– Captura
– Curadoria
– Análise

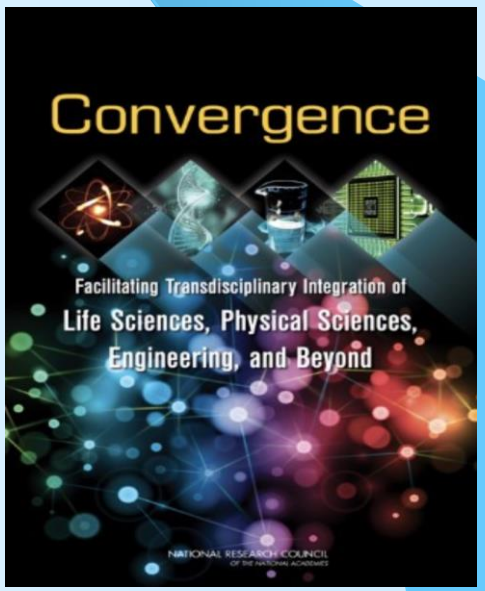
- Dados capturados por instrumentos ou gerados por simulação.
- Dados processados por software.
- Informação/conhecimento armazenados em computadores (em grande escala).
- Pesquisadores analisam arquivos/bases de dados por meio de gerenciamento de dados e estatísticas.

Slides adaptados da apresentação de Jim Gray: **eScience Talk at NRC-CSTB meeting**. http://research.microsoft.com/~gray/talks

Ciência Convergente

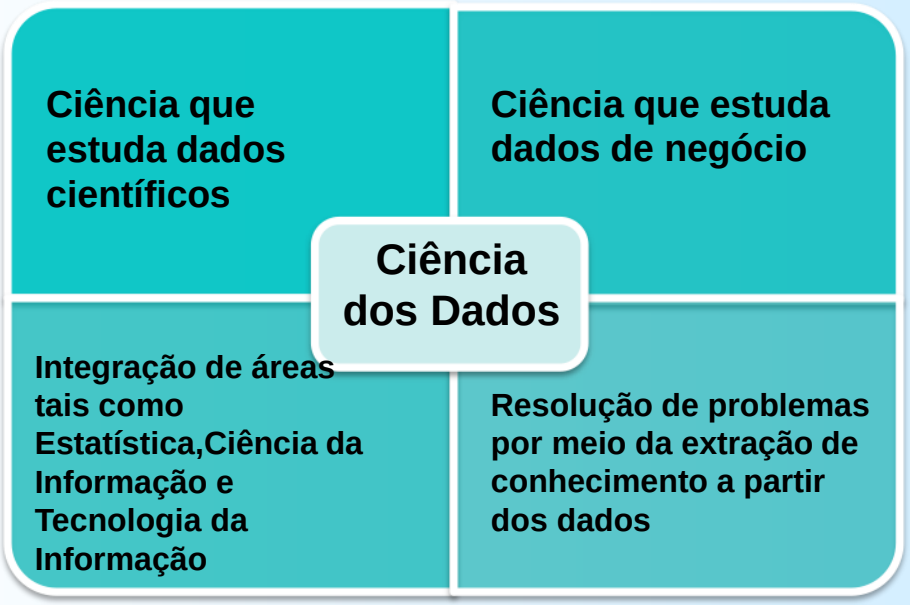
- “the merging of ideas, approaches and technologies from widely diverse fields of knowledge to stimulate innovation and discovery”

"A fusão de idéias, abordagens e tecnologias de campos de conhecimento amplamente diversificados para estimular a inovação e a descoberta"



1. Introdução – Ciência dos Dados

- Em busca por uma definição formal sobre Ciência dos Dados, encontramos diversos trabalhos na literatura
  - Embora muito se discuta sobre a composição das atividades de Ciência dos Dados, o seu conceito ainda não é algo fundamentalmente estabelecido
- Para Zhu e Xiong (2015), há quatro vertentes (perspectivas) que buscam caracterizar Ciência dos Dados

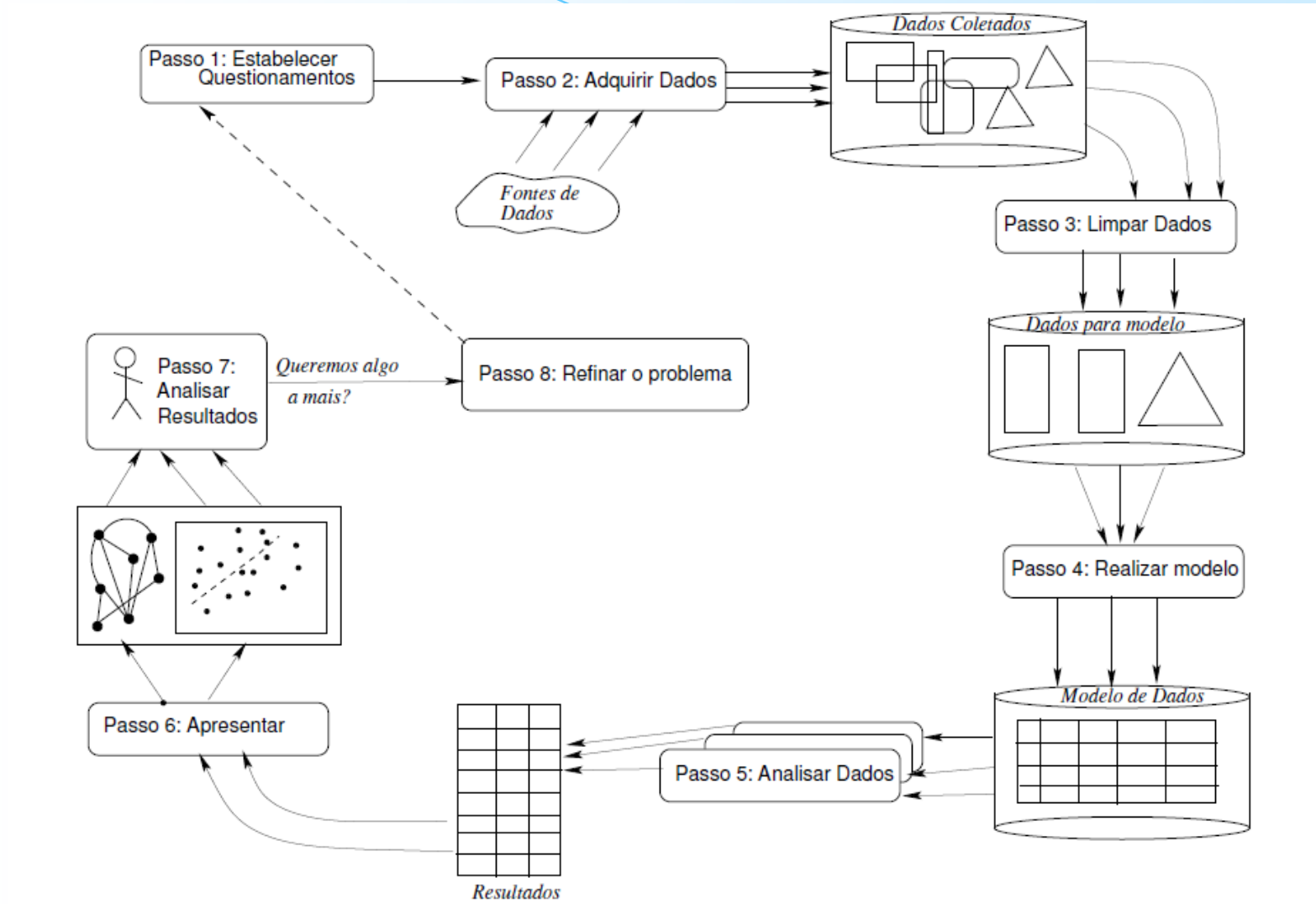


1. Introdução – Ciência dos Dados

- Embora não haja consenso sobre a definição, encontramos como elemento comum em todas as propostas um processo de manipulação, processamento e análise de dados, que visa a descoberta de novos conhecimentos

- Para Alex Dehktyar (2016),
  - Ciência dos dados é uma disciplina que permite tratar o o ciclo de trabalho com os dados, considerando atividades que compreendem desde a aquisição dos dados, passando pela análise dos dados, até o processo de apresentação dos dados e obtenção de novos conhecimentos

1. Introdução – Ciência dos Dados



Cortesia: Alex Dehktyar

2. Questões sobre a Gestão de Dados Científicos

2. Por que a gestão de dados?

PORQUE APLICAR AS TÉCNICAS E CONCEITOS DE GESTÃO DE DADOS?

2. Porque a gestão de dados?

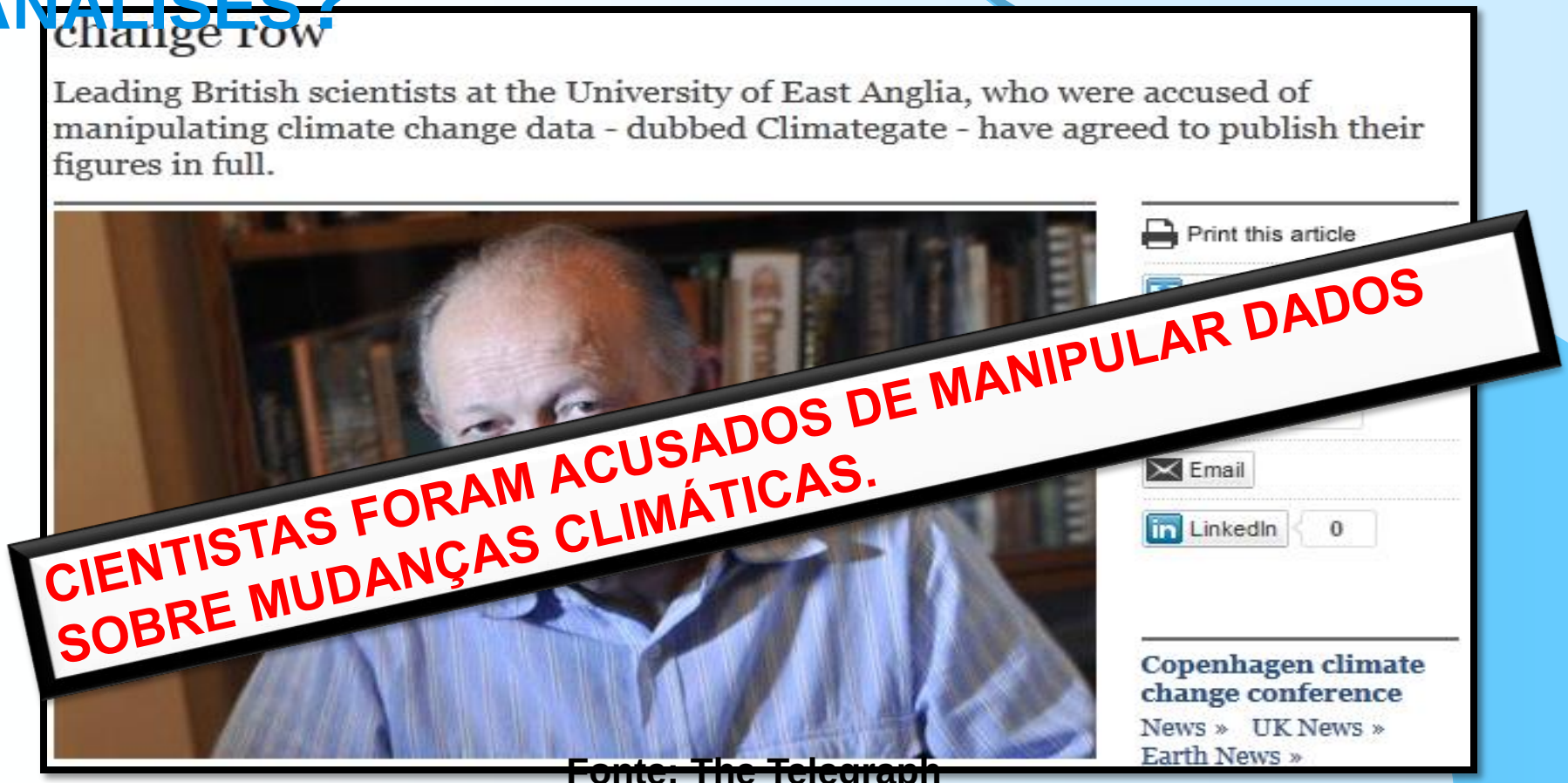
Se seus dados cairem em mãos erradas?

INFORMAÇÕES COM POTENCIAIS IMPLICAÇÕES PARA A SEGURANÇA NACIONAL VAZARAM DO GOVERNO BRITÂNICO.

Fonte: http://news.bbc.co.uk/1/hi/uk/8332445.stm

2. Por que a gestão de dados?

Se for necessária Reproduzir as Análises?



Fonte: The Telegraph

2. Por que a gestão de dados?

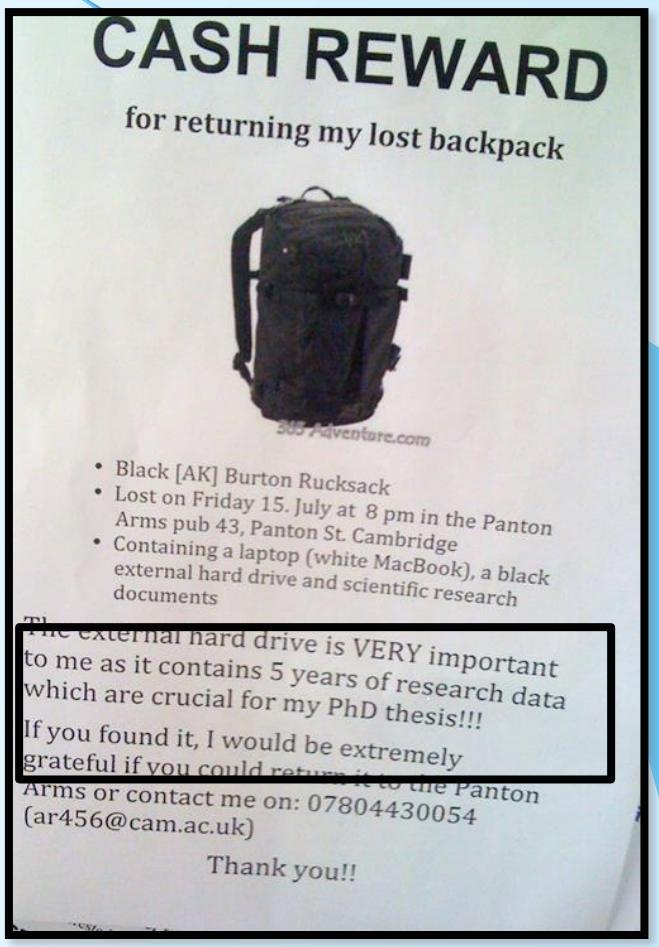
Se este for o seu Instituto de Pesquisa?

Fontes: http://g1.globo.com/sao-paulo/noticia/2010/05/incendio-no-instituto-butantan-destroi-maior-acervo-de-cobras-do-pais.html
https://g1.globo.com/rj/rio-de-janeiro/noticia/2018/09/02/incendio-atinge-a-quinta-da-boa-vista-rio.ghtml

2. Por que a gestão de dados?

SE ESTA FOR A SUA MOCHILA?

"O HD externo é muito importante, pois contém 5 anos de dados de pesquisas…"



Fonte: http://blogs.ch.cam.ac.uk/pmr/2011/08/01/why-you-need-a-data-management-plan

2. Gestão de dados

“Gestão de Dados é a disciplina responsável por definir, planejar, implantar e executar: estratégias, procedimentos e práticas necessárias para gerenciar de forma efetiva os recursos de dados e informações das organizações, incluindo planos para sua definição, padronização, organização, proteção e utilização.”

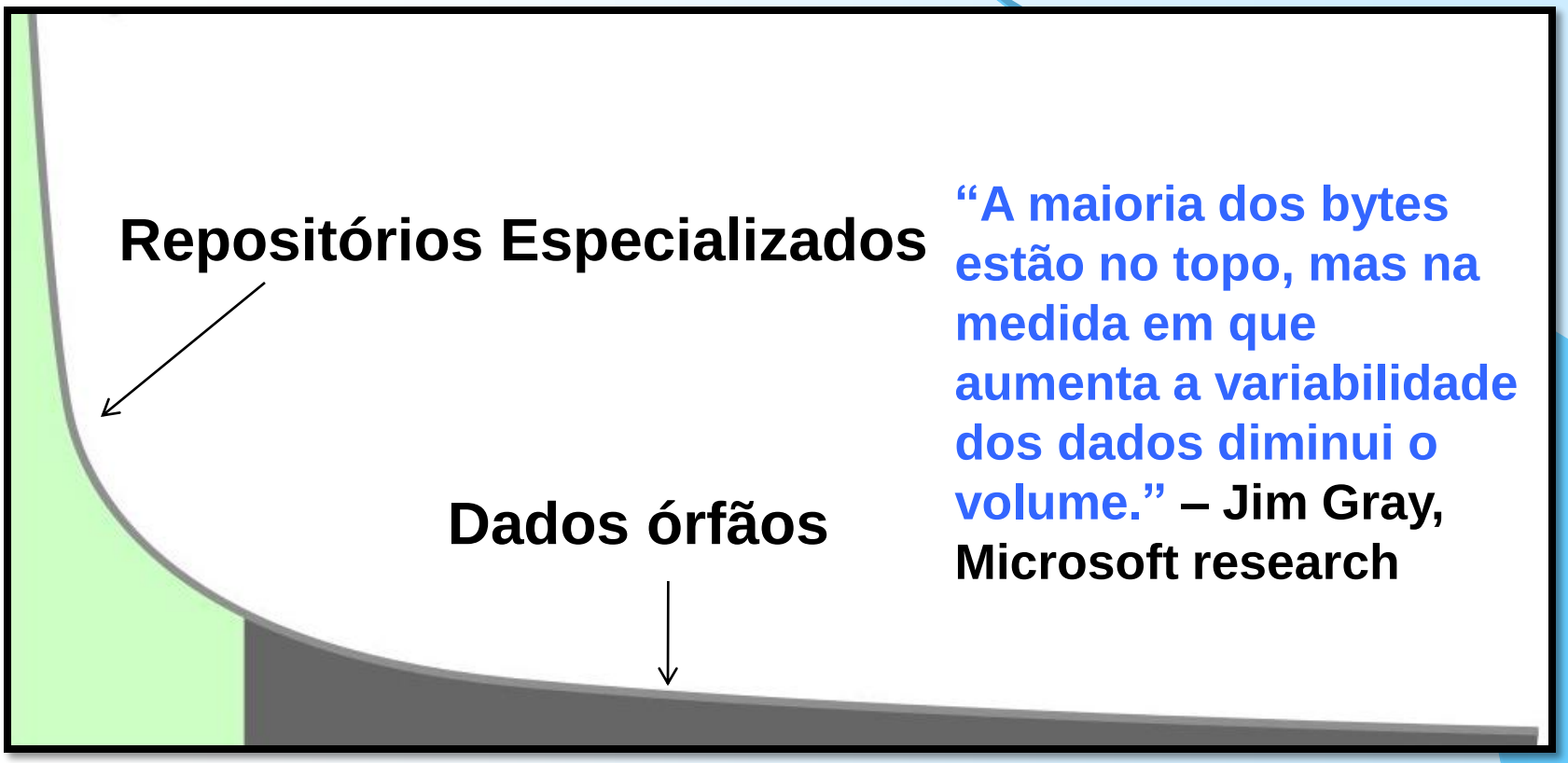
Fonte: DAMA-DMBOK

A Gestão de Dados é um conceito bastante amplo, ela atua nos níveis: Operacional, Gerencial (Tática) e Estratégico.

2. Desafios: Dados órfãos

- Informação que se tornou irrecuperável por estar localizada em dispositivos não mais acessíveis, como notebooks, e que nunca foram transferidas para servidores computacionais;
- Informações perdidas após o desligamento de pesquisadores/funcionários da instituição;
- Dados de Pesquisadores não associados a nenhuma rede

2. Desafios: "The Long tail" da Gestão dos Dados

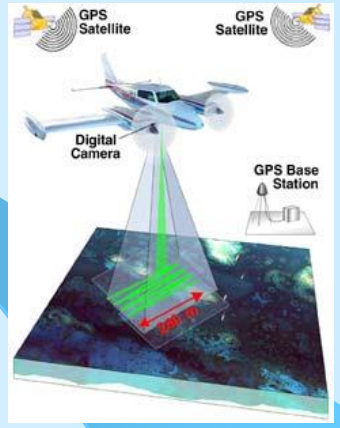


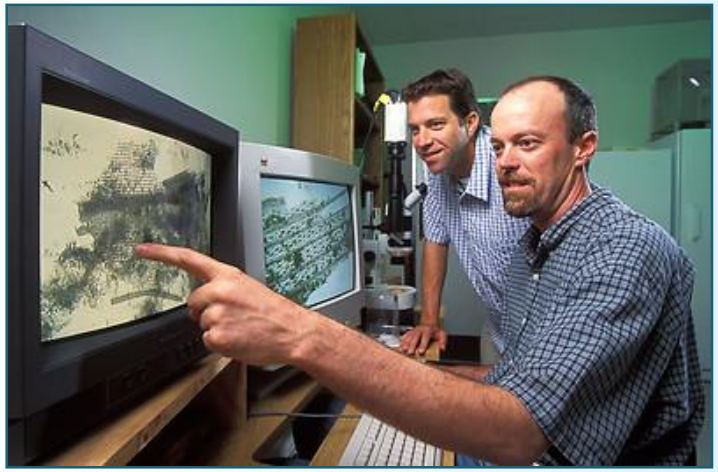
"A maioria dos bytes estão no topo, mas na medida em que aumenta a variabilidade dos dados diminui o volume." – Jim Gray, Microsoft research

Fonte: DataONE

2.Desafios: “Dilúvio” dos Dados

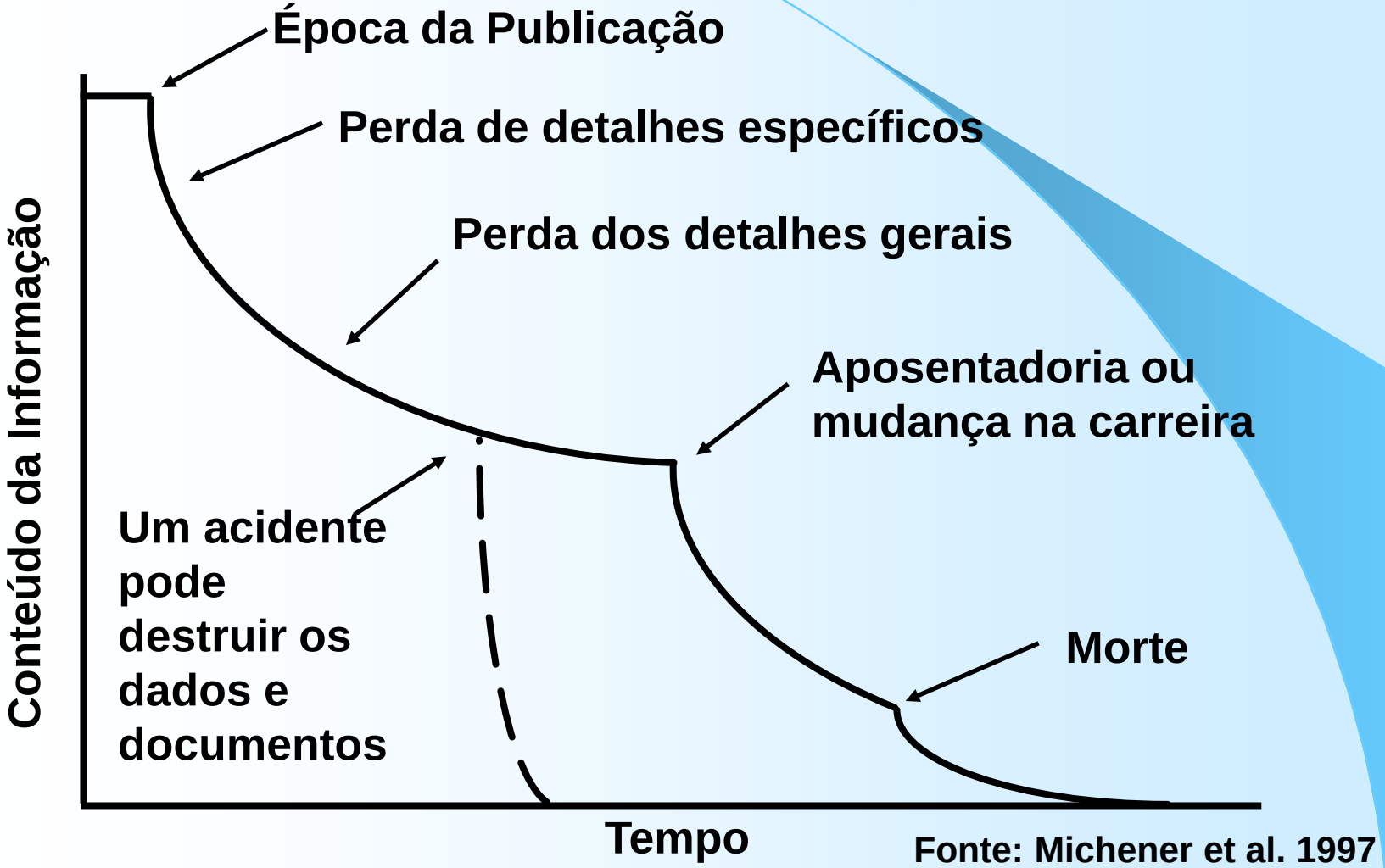
Redes, Sensores, Sensoriamento Remoto, Experimentos, Coletas…





Fonte: www.carboafrica.net

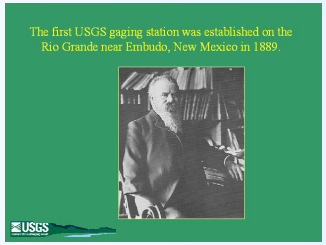
2.Desafios: Disponibilidade dos Dados Científicos



Fonte: Michener et al. 1997

Visão histórica sobre Gestão de Dados

John Wesley Powell fez a primeira medida sistemática do fluxo do Rio Grande Novo Mexico em 1889.



Este processo prematuro de gerenciamento de dados completou 129, transformou num Sistema da USGS chamado National Water Information System (NWIS)

Durante esse período, houveram desafios sobre práticas de gestão de dados, pois …

…a ciências tem focado principalmnte na síntese, interpretação e conclusões, compartilhadas em publicações (normal)

Considerações sobre preservação dos Dados Científicos

“The odds of finding the original data for …papers fell by 17 percent every year after publication.”

“The data are thus unavailable for future researchers to check old results or use for entirely new purposes.”

Current Biology 24, 94–97, January 6, 2014 ©2014 Elsevier Ltd All rights reserved http://dx.doi.org/10.1016/j.cub.2013.11.014

Report

The Availability of Research Data Declines Rapidly with Article Age

Timothy H. Vines,[1,2,*] Arianne Y.K. Albert,[2] Rose L. Andrew,[1] Florence Débarre,[1,4] Dan G. Bock,[1] Michelle T. Franklin,[1,5] Kimberly J. Gilbert,[1] Jean-Sébastien Moore,[1,6] Sébastien Renaut,[1] and Diana J. Rennison[1]
[1]Biodiversity Research Centre, University of British Columbia, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada
[2]Molecular Ecology Editorial Office, 6270 University Boulevard, Vancouver, BC V6T 1Z4, Canada
[3]Women's Health Research Institute, 4500 Oak Street, Vancouver, BC V6H 3N1, Canada
[4]Centre for Ecology & Conservation Biosciences, University of Exeter, Cornwall Campus, Tremough, Penryn TR10 9EZ, UK
[5]Institute for Sustainable Horticulture, Kwantlen Polytechnic University, 12666 72nd Avenue, Surrey, BC V3W 2M8, Canada
[6]Department of Biology, Université Laval, 1030 Avenue de la Médecine, Laval, QC G1V 0A6, Canada

Summary

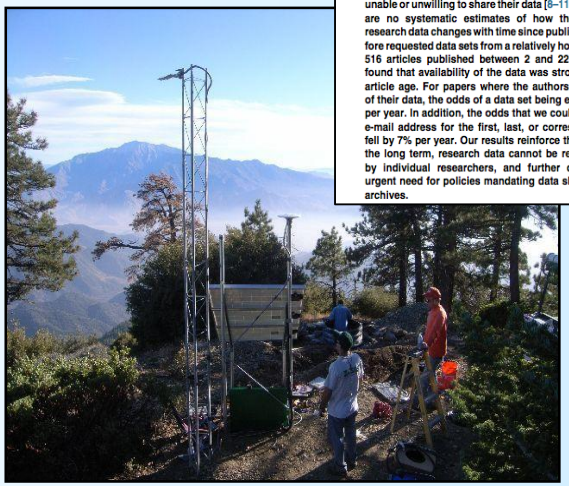
Policies ensuring that research data are available on public archives are increasingly being implemented at the government [1], funding agency [2–4], and journal [5, 6] level. These policies are predicated on the idea that authors are poor stewards of their data, particularly over the long term [7], and indeed many studies have found that authors are often unable or unwilling to share their data [8–11]. However, there are no systematic estimates of how the availability of research data changes with time since publication. We therefore requested data sets from a relatively homogenous set of 516 articles published between 2 and 22 years ago, and found that availability of the data was strongly affected by article age. For papers where the authors gave the status of their data, the odds of a data set being extant fell by 17% per year. In addition, the odds that we could find a working e-mail address for the first, last, or corresponding author fell by 7% per year. Our results reinforce the notion that, in the long term, research data cannot be reliably preserved by individual researchers, and further demonstrate the urgent need for policies mandating data sharing via public archives.

sets (23%) were confirmed as extant. Table 1 provides a breakdown of the data by year.

We used logistic regression to formally investigate the relationships between the age of the paper and (1) the probability that at least one e-mail appeared to work (i.e., did not generate an error message), (2) the conditional probability of a response given that at least one e-mail appeared to work, (3) the conditional probability of getting a response that indicated the status of the data (data lost, data exist but unwilling to share, or data shared) given that a response was received, and, finally, (4) the conditional probability that the data were extant (either “shared” or “exists but unwilling to share”) given that an informative response was received.

There was a negative relationship between the age of the paper and the probability of finding at least one apparently working e-mail either in the paper or by searching online (odds ratio [OR] = 0.93 [0.90–0.96, 95% confidence interval (CI)], $p < 0.00001$). The odds ratio suggests that for every year since publication, the odds of finding at least one apparently working e-mail decreased by 7% (Figure 1A). Since we searched for e-mails in both the paper and online, four factors contribute to the probability of finding a working e-mail: (1) the number of e-mails in the paper and (2) the chance that any of those worked and (3) the number of e-mails we could find by searching online and (4) the chance that any of those worked. The total number of e-mail addresses we found in the paper decreased with age (Poisson regression coefficient = −0.07, SE = 0.01, $p < 0.0001$) from an average of 1.17 in 2011 to 0.42 in 1991 (Figure 2A), and there was a slight positive effect of article age on the number of e-mails we found online (Poisson regression coefficient = 0.015, SE = 0.007, $p < 0.05$; Figure 2C). Moreover, the chance that an e-mail found in the paper or online appeared to work also showed a relationship with article age (OR = 0.96 [0.926–0.998, 95% CI], $p < 0.05$; and OR = 0.97 [0.936–0.997, 95% CI], $p < 0.05$; respectively), such that the odds that an e-mail appeared to work declined by 4% and 3% per year since publication, respectively (Figures 2B and 2D).

We note that eight e-mail addresses generated an error message but did lead to a response from the authors. It also seems likely that some addresses failed but did not generate



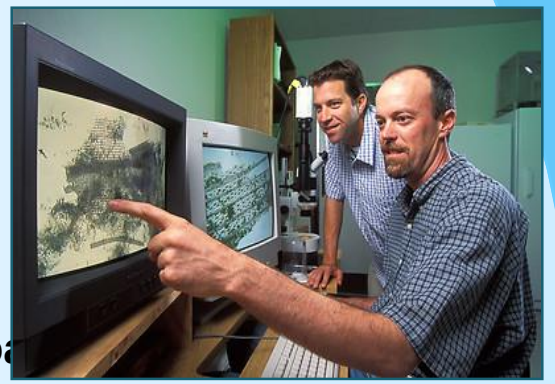
Fonte:
http://www.sciencedirect.com/science/article/pii/S0960982213014000
Photo Source:
EROS Data Center, USGS Archivist

Princípios para Gestão de Dados Científicos

Princípios para gestão de dados científicos:

- **Os dados são ativos de ciência incrivelmente valiosos e fundamentais.**
- **Os dados coletados hoje são *snapshots*. Eles não estarão disponíveis amanhã da mesma forma;**
- **Precisamos administrar e proteger os dados científicos que coletamos, de modo que os dados sejam acessíveis, compreendidos, reproduzíveis e reutilizáveis**

Objetivo: mudar a cultura dos cientistas, de modo que os dados sejam gerenciados efetivamente como parte natural do fluxo de trabalho da pesquisa.



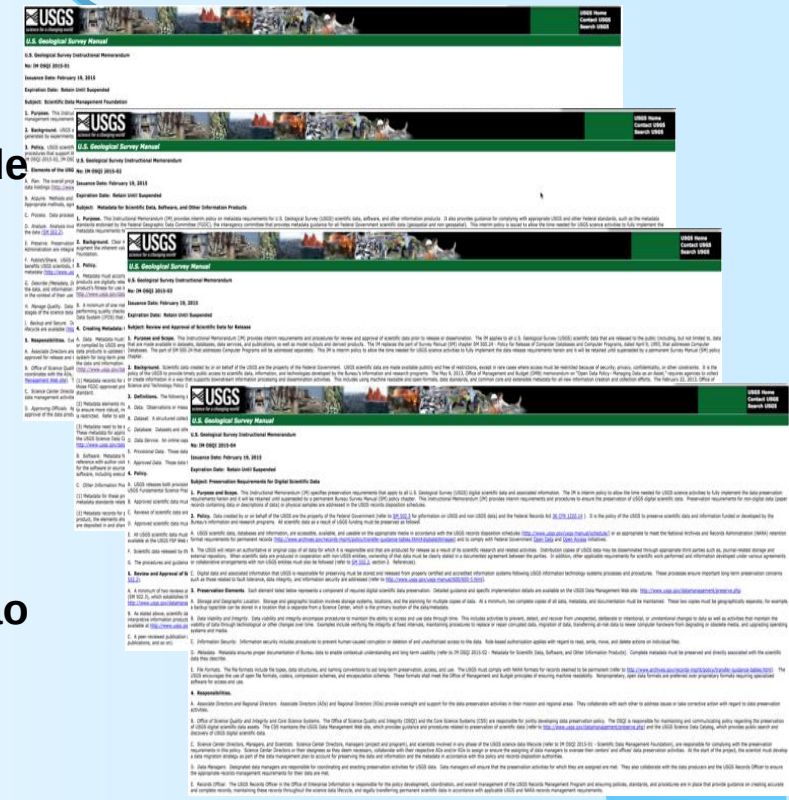
Desafio Institucional:

Definição de uma Política de Dados

Exemplo da USGS:

Política de Gestão de Dados Científicos

- Estabelece os fundamentos para Gerenciamento de Dados Científicos
- Define ferramentasde Software, Padrões de Metadados para Dados Científicos
- Estabelece um processo para revisão e aprovação de publicação de Dados Científicos
- Define os **Requisitos** para a preservação de Dados Científicos Digitais



Políticas foram introduzidas em 2015 de maneira sistemática na USGS

https://www2.usgs.gov/fsp/policies.asp

Desafio Institucional:
Implementação de Gestão de Dados Científicos
Exemplo da USGS:

USGS Public Access Plan

- Regarding science data, the Plan introduces a “new normal”:
  - USGS scientists must release the data upon which their scientific publications are based.

- A detailed data management plan for all research projects
- Data in machine readable, open formats
- Standard metadata describing the data
- A digital object identifier for the data, recorded in the metadata record
- Approvals: requiring data review, metadata review, and either Center Director or Bureau Approving Officials
- Hosted and shared from a reliable, repository location
- Metadata shared through the USGS Science Data Catalog

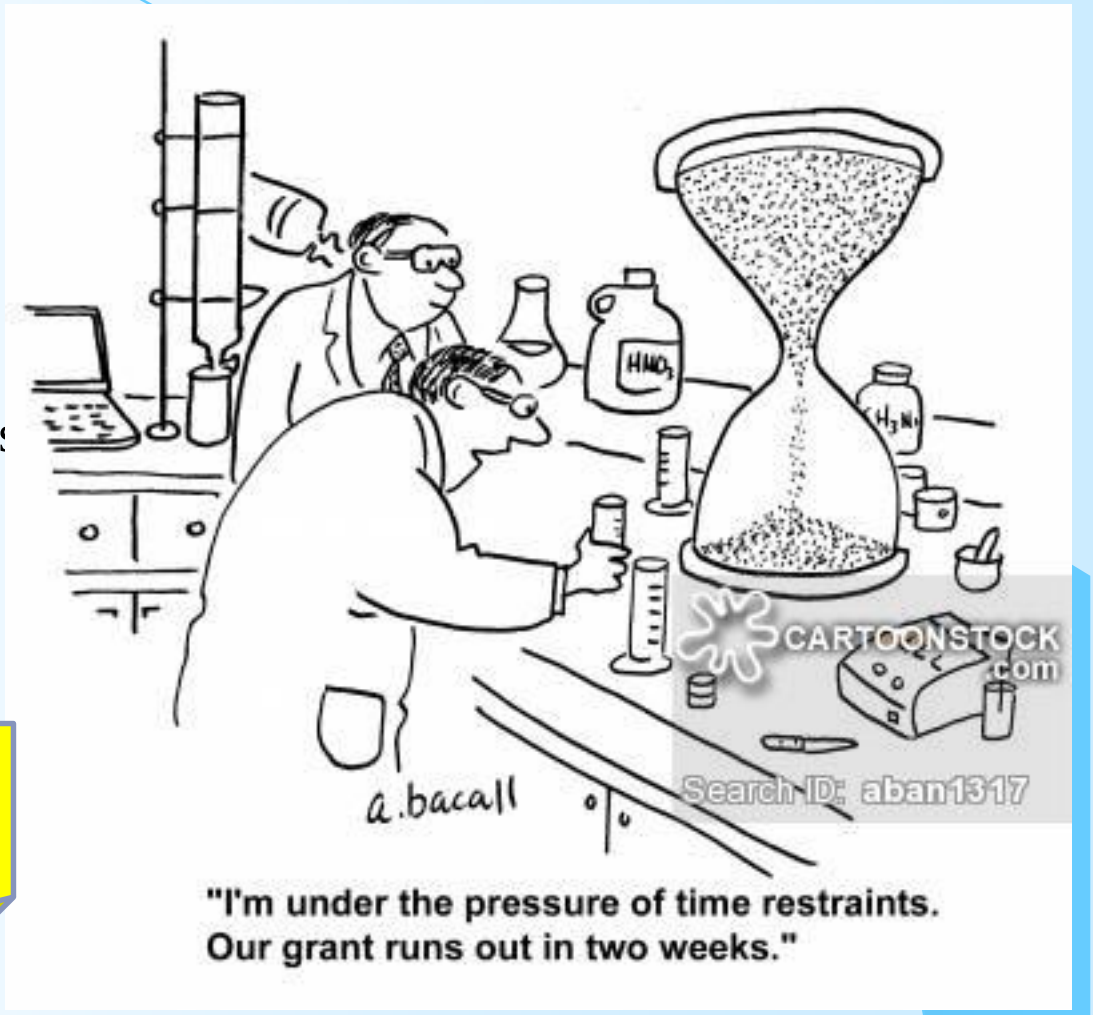
“Public Access to Results of Federally Funded Research at the U.S. Geological Survey: Scholarly Publications and Digital Data”

http://www.usgs.gov/quality_integrity/open_access/

Pressões dos Cientistas

- Publicação
- *Deadlines*
- Políticas
- Recursos
- Diminuição na equipe
- Congelamento nas contratações
- Resistência em Colaborações
- Muitas prestações de contas
- …

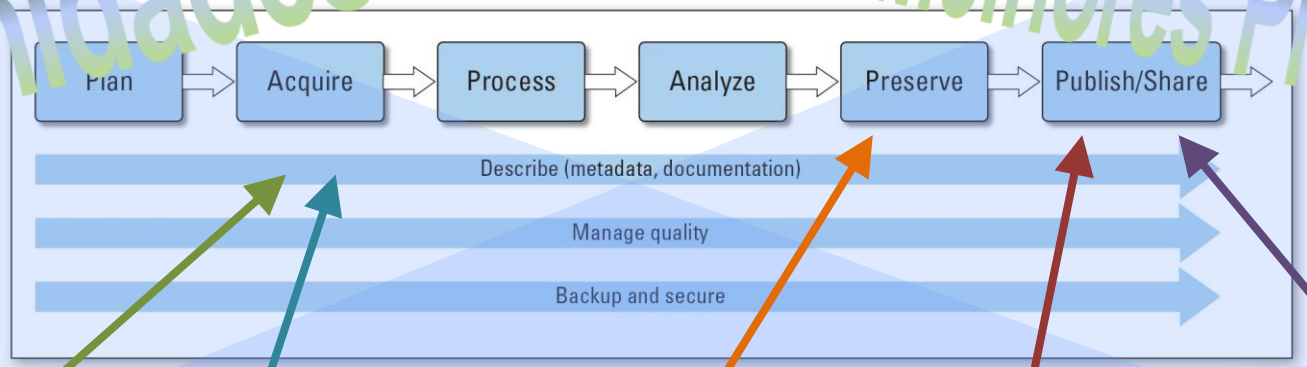
ENTÃO, OS CIENTISTAS FICARÃO FELIZES EM SABER QUE TERÃO QUE GERAR METADADOS ?????

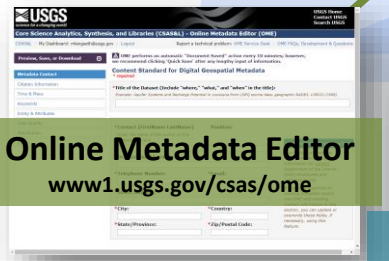


Definição de um modelo de Ciclo de Vida de Dados Institucional. Exemplo da USGS
Ciclo de Vida dos Dados Científicos – Ferramentas & Serviços

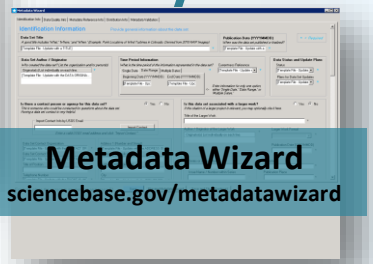
Comunidades

Melhores Práticas





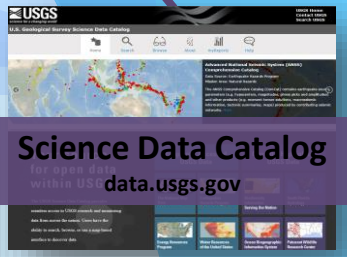
Online Metadata Editor
www1.usgs.gov/csas/ome



Metadata Wizard
sciencebase.gov/metadatawizard

DOI Creation Tool
www1.usgs.gov/csas/doi

ScienceBase
sciencebase.gov



Science Data Catalog
data.usgs.gov

Software – Próxima fronteira

Similar ao processo de publicação dos Dados
O software que Analisa os dados precisa acompanhar os artigos
Considerações:

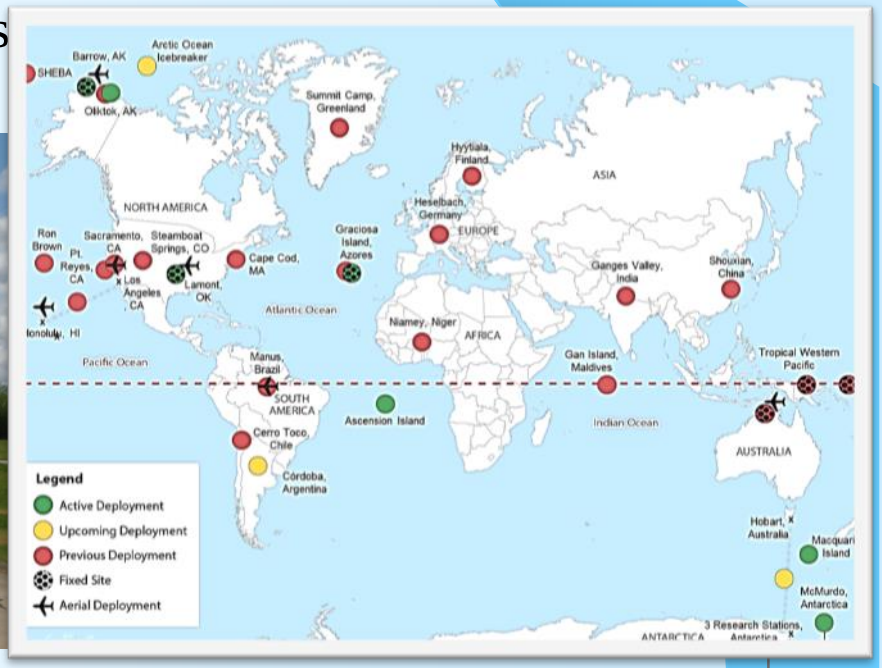
- Categorias de disponibilização – Informal e formal
- Formal: Disclaimers (Provisório/Aprovado) https://www2.usgs.gov/fsp/fsp_disclaimers.asp
- Licenses – Código pode ser de domínio público e/ou incluir restrições de terceiros;
- Estratégias de Documentação - https://github.com/usgs/best-practices
- Code Reviews (PII/Security, scientific verification, standards)- https://github.com/usgs/best-practices
- Obtenção de um DOI https://github.com/usgs/best-practices/blob/master/doi.md

Exemplo de Projetos

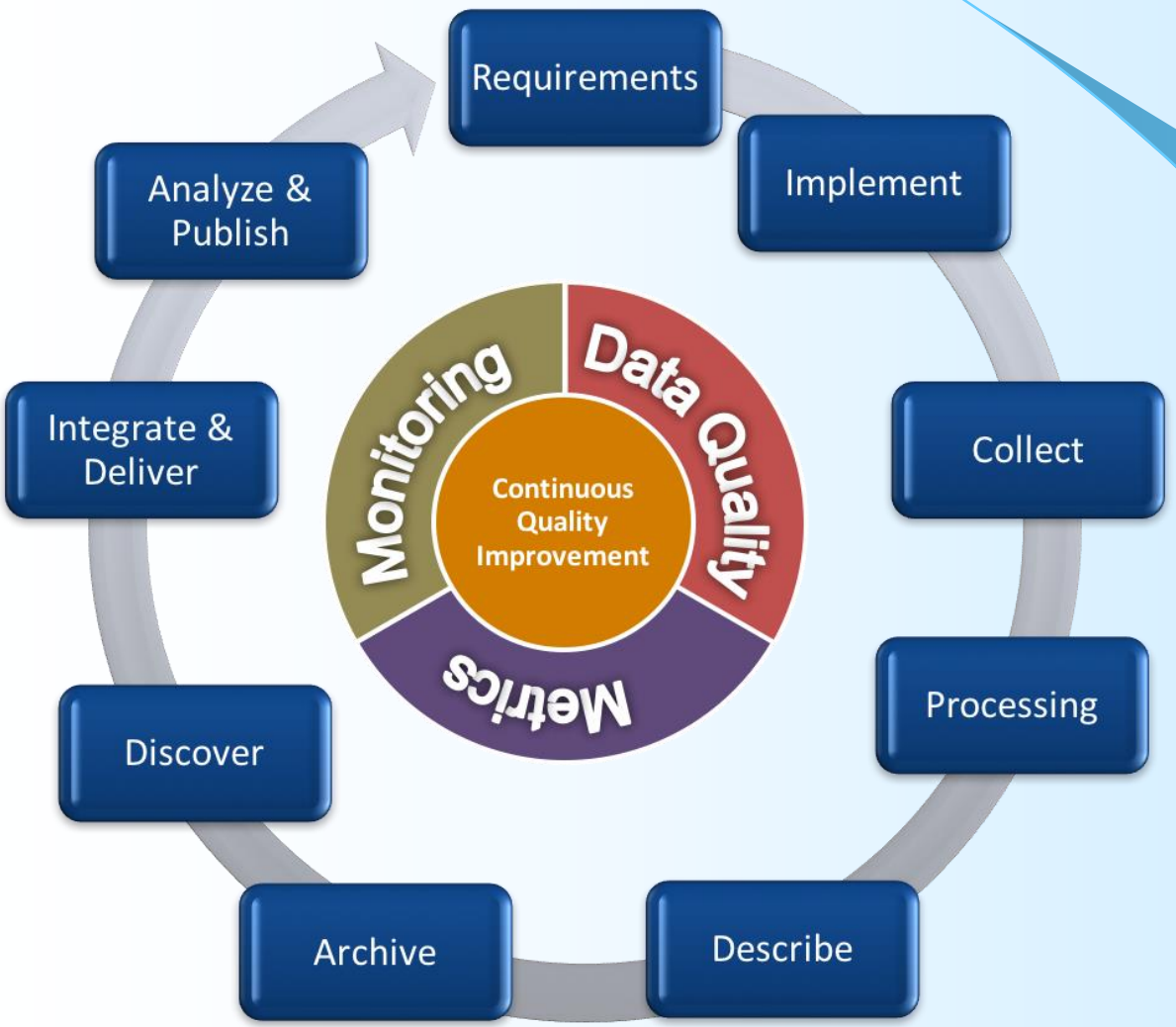
The Atmospheric Radiation Measurement (ARM) Facility
Data and Computing Management (DoE/USA)

Objetivo do ARM:

fornecer uma detalhada e precisa descrição da atmosfera da terra em diversos regimes climáticos para resolver as incertezas no clima e nos modelos dos sistemas terrestres que direcionam o desenvolvimento de soluções sustentáveis para a Energia do País &desafios ambientais.

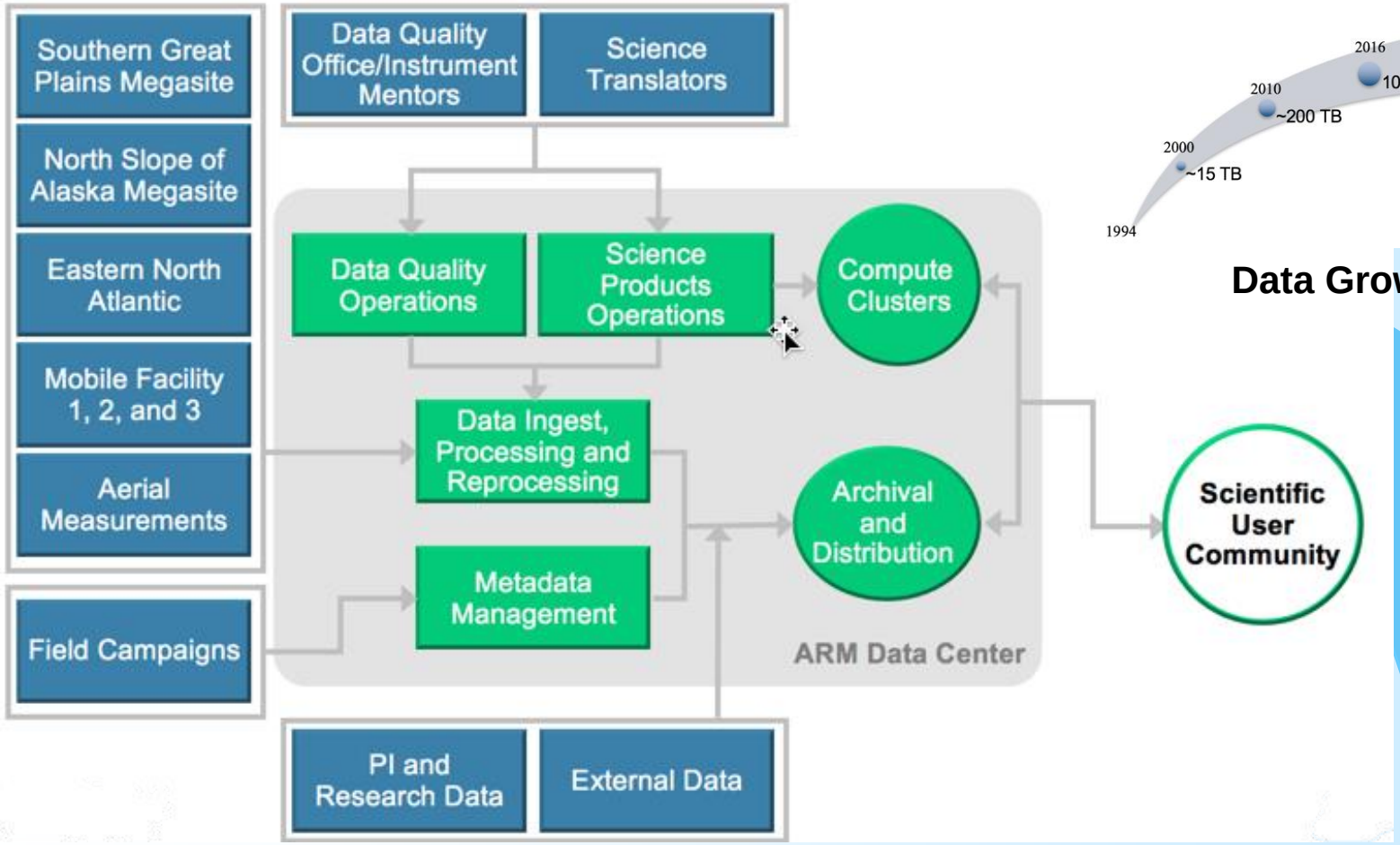


Ciclo de Vida dos Dados do ARM



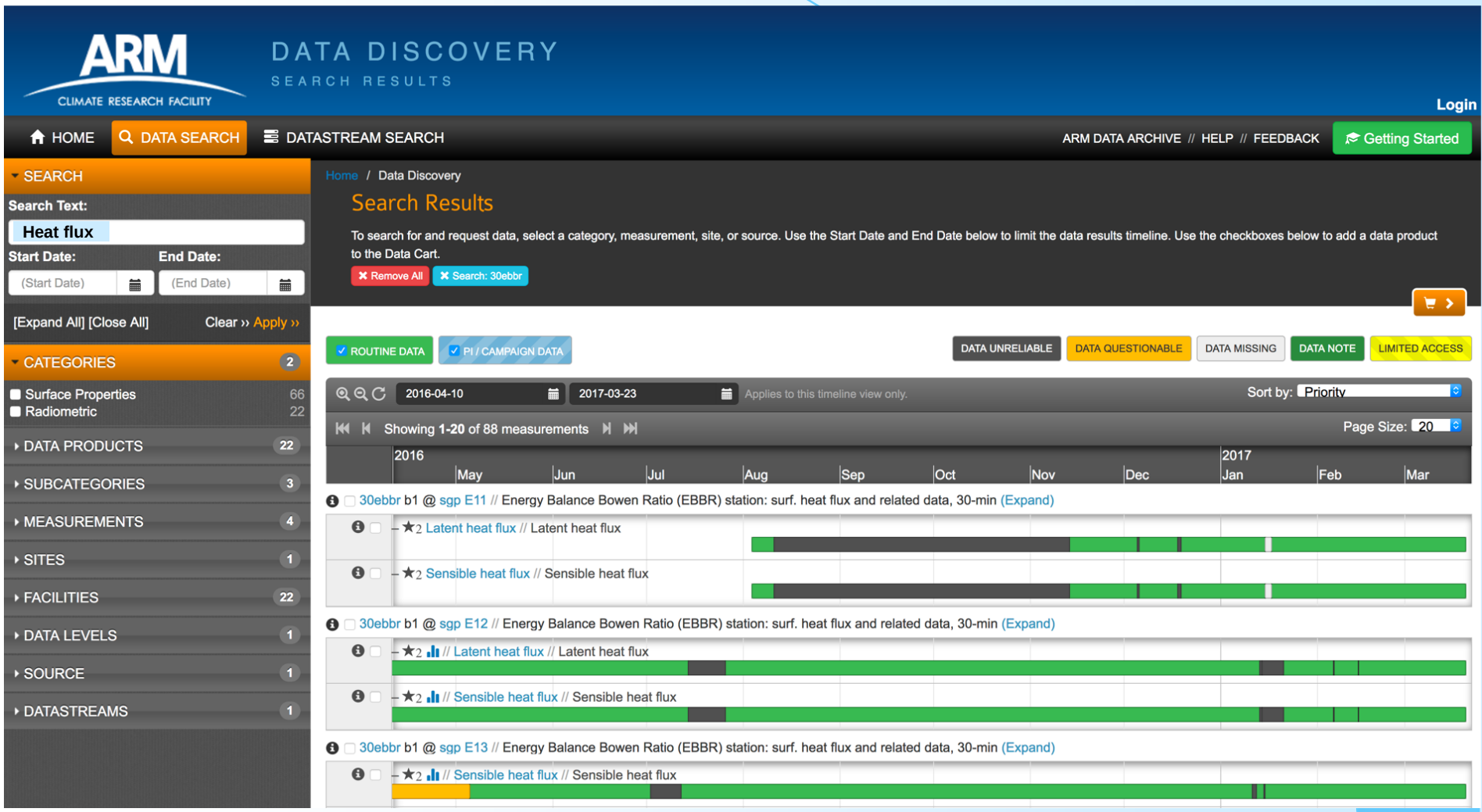
- **Matured processes are used in every component of the Data Lifecycle**
- **Continuous integration helps improve data quality**
- **A variety of tools help monitor data lifecycle components 24/7**

Fluxo de Dados do ARM– Visão Geral



ARM Permanent Sites provide Long-Term Data.
Mobile Sites and Aircraft Increase Diversity.

Ferramenta para Acesso aos Dados



ARM - Software Stack

Operation System: RHEL 7

Compilers:

- **Intel, PGI, GCC**

Libraries:

- **MPI**
- **OpenMPI**
- **CUDA**
- **NetCDF, HDF5**
- –

Development tools:

- **Intel IDE, debuggers**
- **GDB**
- **Valgrind**
- **Git, Mercurial**

Software (not a complete list):

- **Python**
- **PyART***
- **ADI***
- **R**
- **NCL, NCO**
- **Ferret**
- **MATLAB with Image Toolkit**
- **ARM software (Py-ART, ADI etc.)**
- **IDL**
- **Spark, Cassandra**

Software environment management using Spack/modules

Job Scheduling:

- **Moab and Torque**
- **Two common login nodes per enclave**
- **Single queue, which helps with bursting, Allowed to burst beyond purchased nodes (30% - 50%)**
- **Fairshare algorithm**

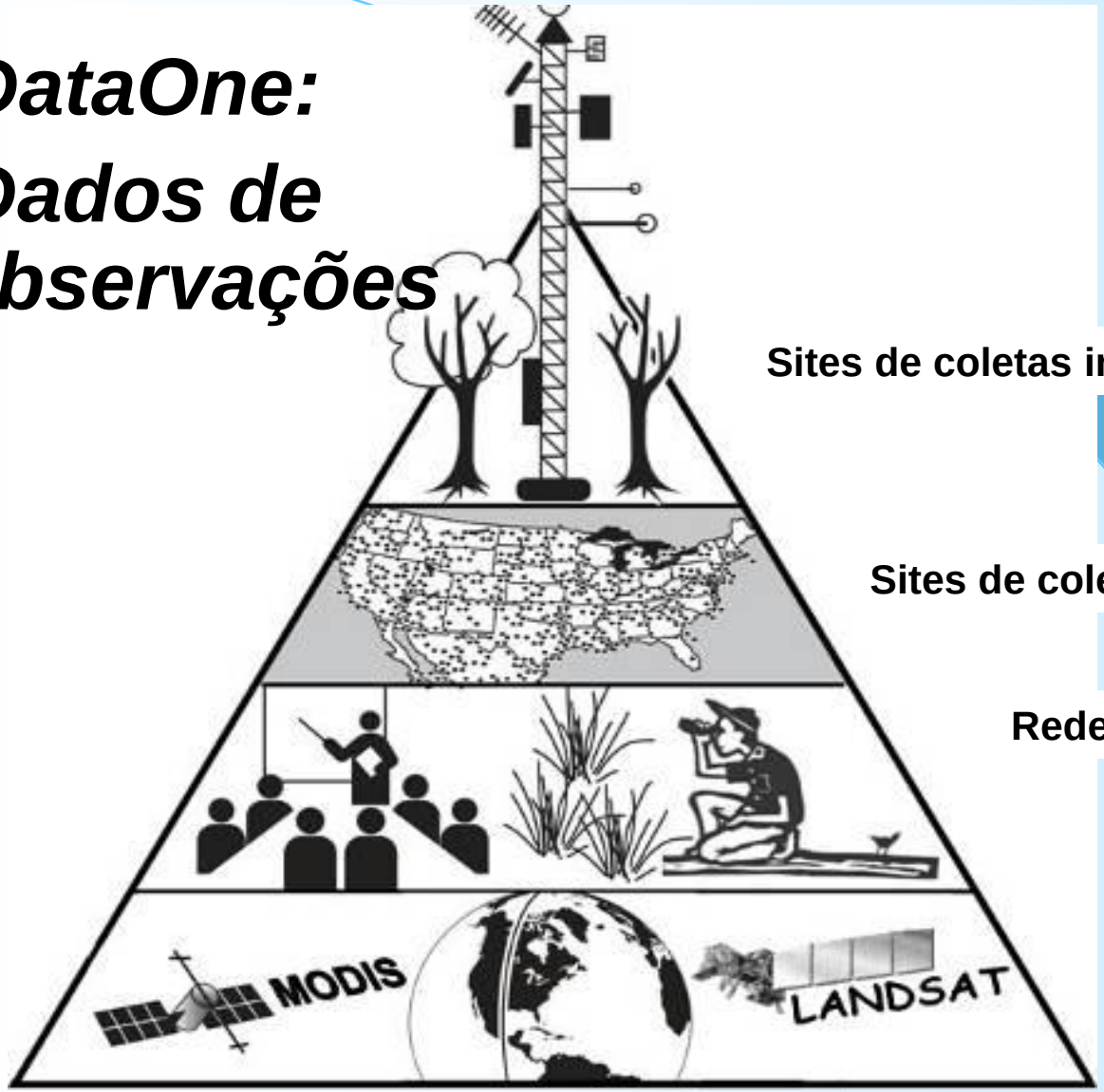
http://adc.arm.gov/tutorials/cluster/stratusclusterquickstart.html#available_software

Providing universal access to data about life on earth and the environment that sustains it.

dataone.org

William Michener, PI (U. New Mexico)
Co-PIs: Suzie Allard, Matt Jones, Dave

DataOne: Dados de Observações



Sites de coletas intesivas

Sites de coletas extensivas

Redes de Voluntários e Educa

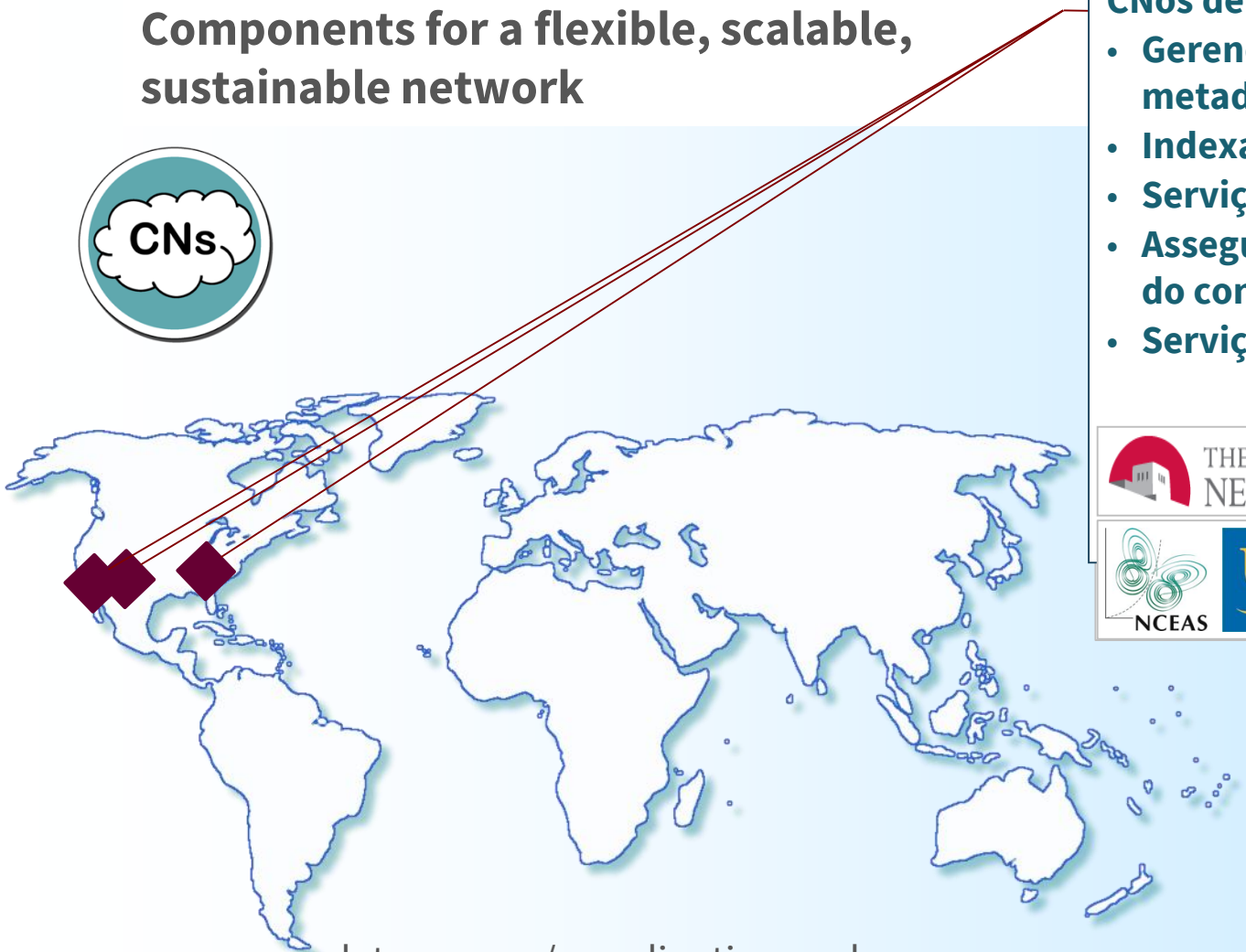
Sensoriamento Remoto

Adapted from CENR-OSTP
Courtesy of DataONE

Coordinating

Federação distribuída de Serviços

Components for a flexible, scalable, sustainable network



CNós de Coordenação

- **Gerencia o catálogo de metadados**
- **Indexação da busca**
- **Serviços de rede**
- **Assegura a disponibilidade do conteúdo (preservação)**
- **Serviços de replicação**

THE UNIVERSITY of NEW MEXICO

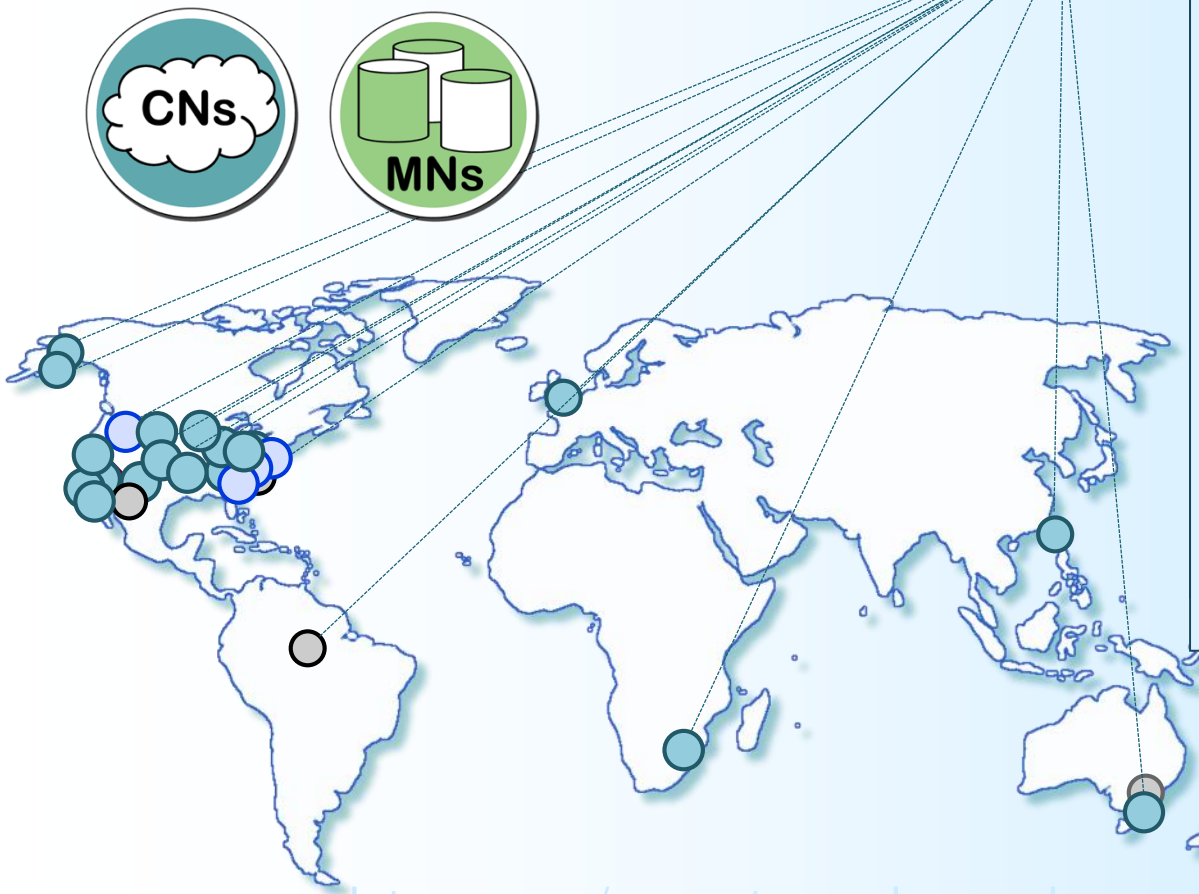
OAK RIDGE National Laboratory

NCEAS

UCSB

www.dataone.org/coordinating-nodes

43 Nós Membros



Membros

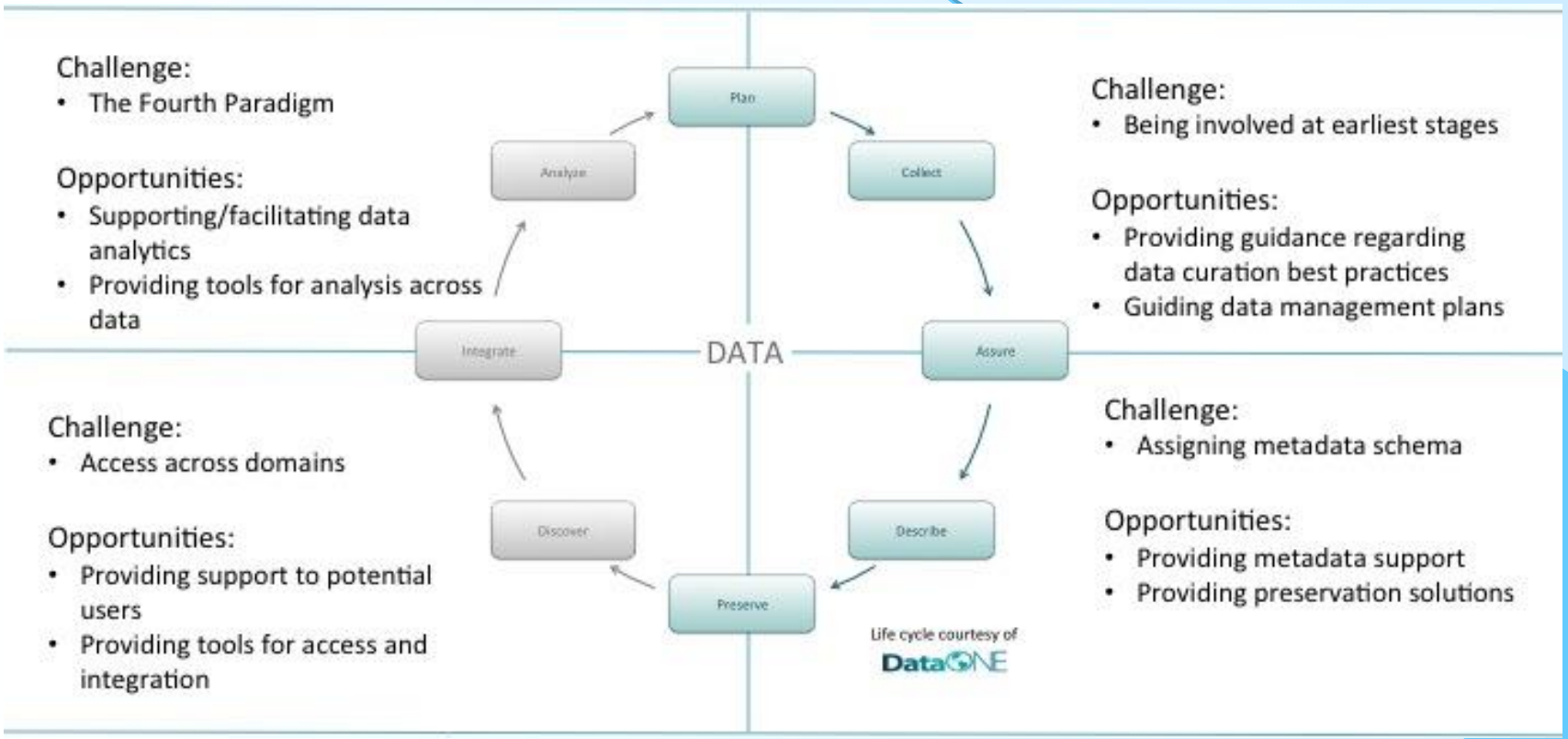
Nós Membros

- **Várias instituições**
- **Serviços para a comunidade local**
- **Gerencia seus dados**
- **Mantem réplcas**

www.dataone.org/current-member-nodes

Solutions for Researchers:
Supporting , Access and Use

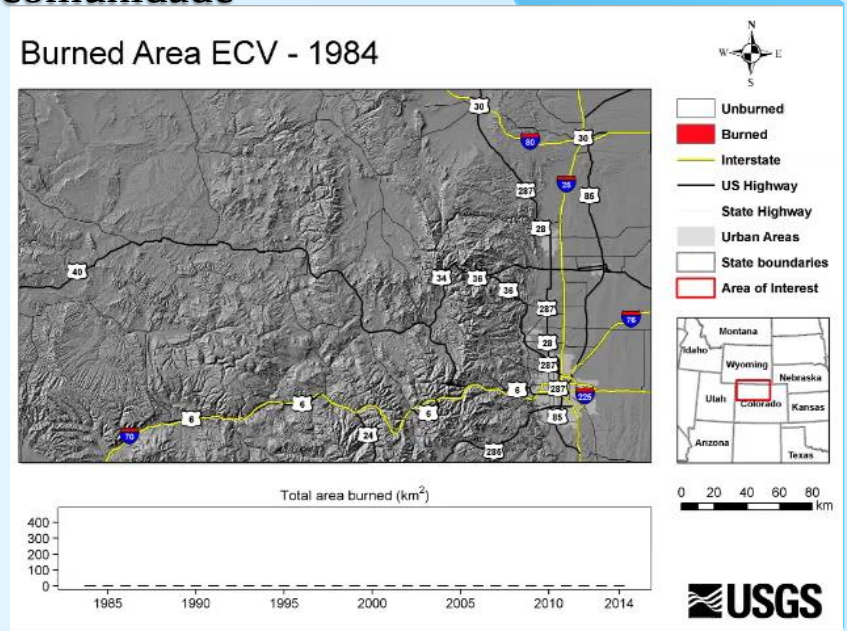
Ciclo de Vida & Desafio a Profissionais da Informação



Source: Allard, S. (2012). The Data Life Cycle & Information Professionals. Third Annual ASIST Research Data Access and Preservation Summit. New Orleans, LA. 21 March 2012.

Conclusão
Lições aprendidas: O que é necessário para Implementar & Sustentabilidade

- Considerar o Ciclo de Vida dos Dados – Gerenciamento em Longo Período e integração de dados
- Ferramentas – Fáceis de usar, treinamento e apoio;
- Formação de recursos humanos – Abordagem interdisciplinar
- Recursos: Modelos de Apoio, Pessoas engajadas, Especialistas
- Políticas – Apoio de Regulamentações Institucionais, Sistema de valorização
- Campanhas de Concientização e engajamento da comunidade
- Métricas – Avaliar os Impactos/Resultados



Trabalho Necessário

COLABORAÇÃO

SUSTENTABILIDADE

Pesquisa & Promoção de:

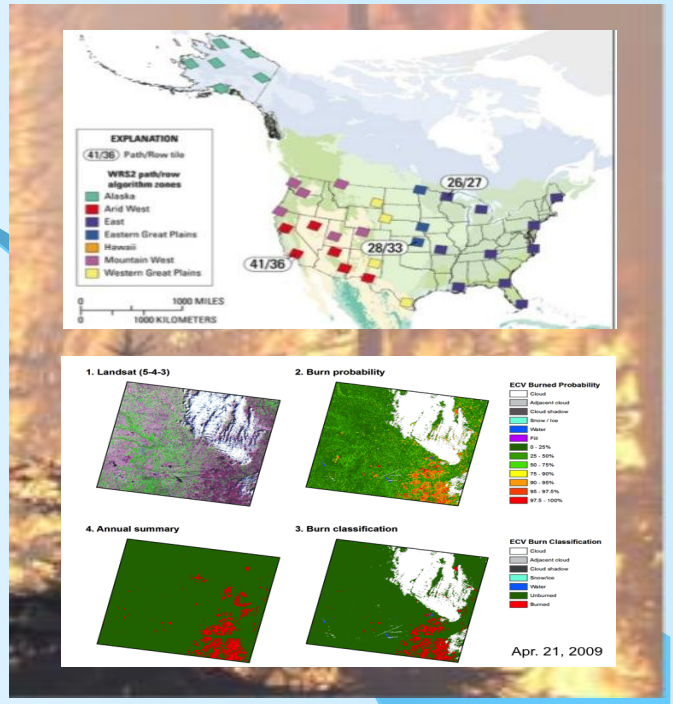
- Princípios Data FAIR:
  - *Findable*
  - *Accessible*
  - *Interoperable*
  - *Re-usable*
- Integração de Dados
- Análise e Visualização de Dados

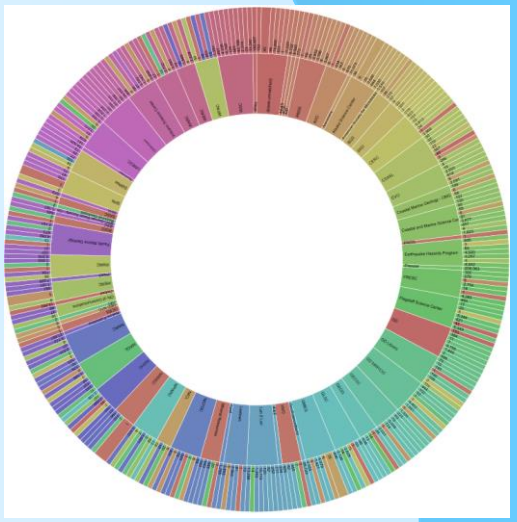
Formação de pessoas:

- Abordagens Interdiciplinares
- Uso e reuso de dados em sala de aula
- Áreas emergentes de estudo

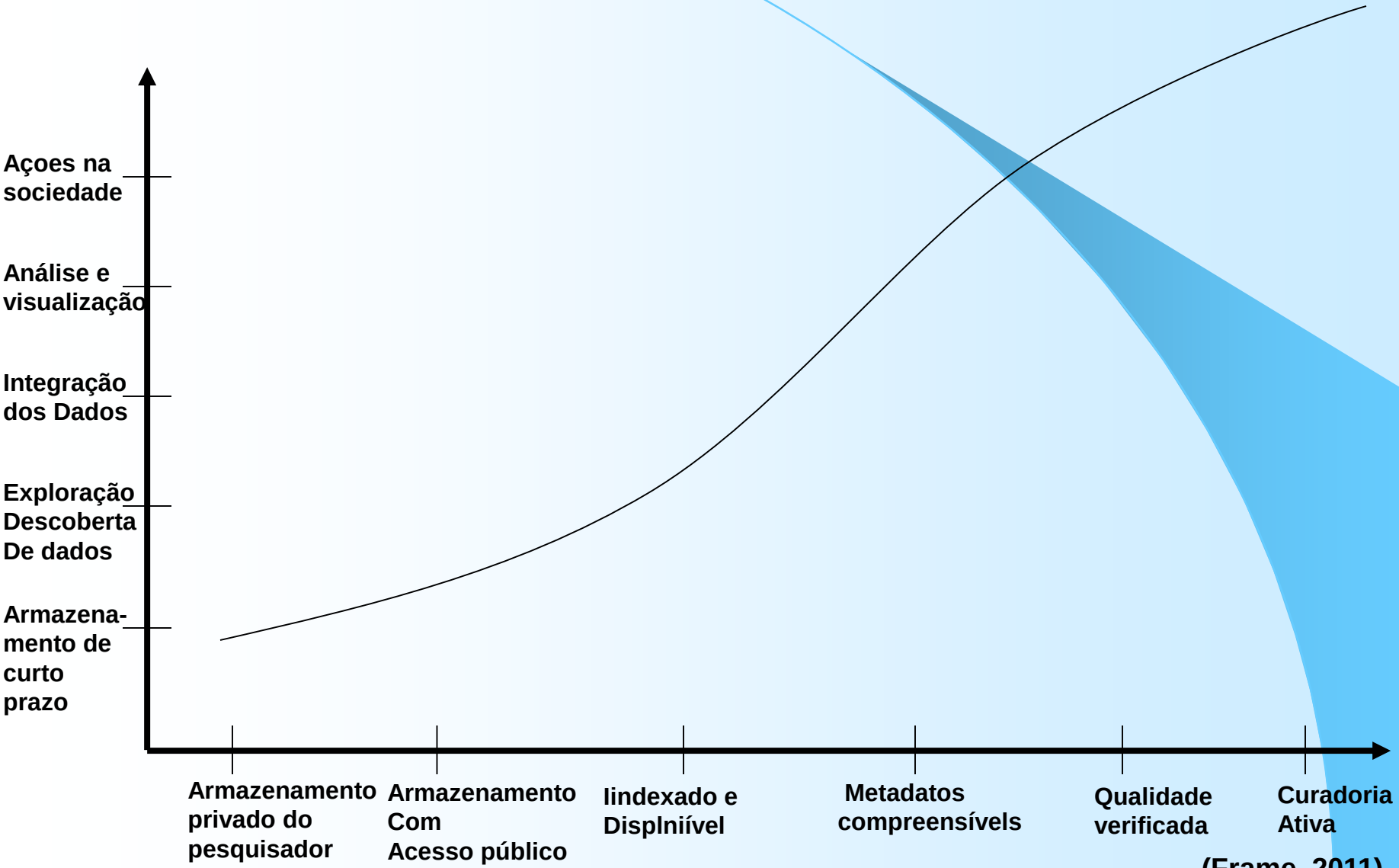
Avanços/Mudanças culturais:

- Sistema de Reconhecimento
- Investimentos em Longo Prazo
- Concientização / Divulgação





Conclusão – Por que nós devemos garantir boas práticas na gestão dos dados?



(Frame, 2011)

Referências

[1] Public, community driven open source repository on GitHub https://github.com/usgs;
[2] USGS hosted repository https://code.usgs.gov; will mirror projects in the USGS organization on GitHub [1]
[3] Disclaimers https://www2.usgs.gov/fsp/fsp_disclaimers.asp;
[3a] Provisional https://www2.usgs.gov/fsp/fsp_disclaimers.asp#11;
[3b] Approved https://www2.usgs.gov/fsp/fsp_disclaimers.asp#5
[4] Public Domain, Creative Commons License https://creativecommons.org/publicdomain/zero/1.0/
[5] Example License file https://github.com/usgs/earthquake-design-ws/blob/master/LICENSE.md
[6] USGS hosted bitbucket, internal with GS domain credentials, and public https://my.usgs.gov/bitbucket/repos?visibility=public
[7] USGS hosted gitlab, internal only, GS domain credentials, https://gitlab.cr.usgs.gov/users/sign_in
[8] USGS technical support teams website https://tst.usgs.gov/
[9] USGS Digital Object Identifier (DOI) information https://github.com/usgs/best-practices/blob/master/doi.md
[10] USGS community best practices at https://github.com/mguy-usgs/best-practices
[11] Suzie Allard (University of Tennesse - UT) DataONE – Slides of III Workshop on Data Science - Escola Politécnica da USP - 2017
[12] Giri Prakash (Oak Ridge National Laboratory - ORNL) ARM program – Slides of III Workshop on Data Science - Escola Politécnica da USP - 2017
[13] Mike Frame (USGS Deputy of Data Science) – USGS Data Management Slides of III Workshop on Data Science - Escola Politécnica da USP - 2017

Referências Adcionais

CLEVELAND, W. S. Data science: an action plan for expanding the technical areas of the field of statistics. International Statistical Review, Blackwell Publishing Ltd, v. 69, n. 1, p. 21-26, 2001. ISSN 1751-5823.

DONOHO, D. 50 years of Data Science. 2015. Disponível em: http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf

PRESS, G. A Very Short History Of Data Science. 2013. http://goo.gl/isvN0hi. Acessado em: 03 set. 2016.

ZHU, Y.; XIONG, Y. Defining Data Science. CoRR, abs/1501.0, 2015

DAMA International, The DAMA Guide to the Data Management Body of Knowledge - DAMA-DMBOK, Technics Publications, LLC, 2010.

Data ONE Best Practices
http://www.dataone.org/best-practices.

FRAME, M. LifeCycle and Metadata. In: IS 590 Environmental Informatics. University of Tennessee. 2015.

Referências adcionais

HEY, T.; TOLLE, K. The fourth paradigm data-intensive scientific discovery. Redmond, Wash.: Microsoft Research, 2009.
http://research.microsoft.com/en-us/UM/redmond/about/collaboration/fourthparadigm/4th_PARADIGM_BOOK_complete_HR.pdf

LAUDON K.C. & LAUDON J.P. Management Information Systems, Capítulo 1. 13º. Ed., Perason, 2013.

WIGGINS, A. et al. Data Management Guide for Public Participation in Scientific Research. DataONE Public Participation in Scientific Reserarch Working Group, 2013.
http://www.dataone.org/sites/all/documents/DataONE-PPSR-DataManagementGuide.pdf
STRASSER, C. et al. Primer on Data Management: What you always wanted to know. 2012. California Digital Library, 2013.
http://escholarship.org/uc/item/7tf5q7n3

BRASIL. Padrões de Interoperabilidade de Governo Eletrônico - e-PING. Comitê Executivo de Governo Eletrônico. Versão 2013.
http://eping.governoeletronico.gov.br/.

CORRÊA, P.L.P. ARQUITETURA PARA INTEGRAÇÃO DE SISTEMAS DE INFORMAÇÃO E BANCO DE DADOS DE BIODIVERSIDADE DO MINISTÉRIO DO MEIO AMBIENTE, MMA. 2012.

WIDaT 2018

II WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

Universidade Federal da Paraíba (UFPB)

Data Science tendências e desafios

Prof. Dr. Pedro Luiz Pizzigatti Corrêa

Grupo de Estudo, Pesquisa e Extensão em Big Data

Escola Politécnica da USP

pedro.correa@usp.br

wds.poli.usp.br