Conceitos de Ciência dos Dados e Big Data

PCS5787 – Ciência dos Dados
Prof. Dr. Pedro Luiz Pizzigatti Corrêa
24 de Setembro de 2020

Agenda

- Introdução
  - O conceito de Dado x Informação x Conhecimento
  - Big Data, Ciência dos Dados, Gestão dos Dados
- Gestão dos dados – Importância e Desafios
- O modelo de dados do ponto de vista do *Big Data*
- Desafios relacionados ao *Big Data* e Ciência dos Dados
- Exercícios

Introdução

O conceito de Dado x Informação x Conhecimento

Conceito Dados, Informação e Conhecimento

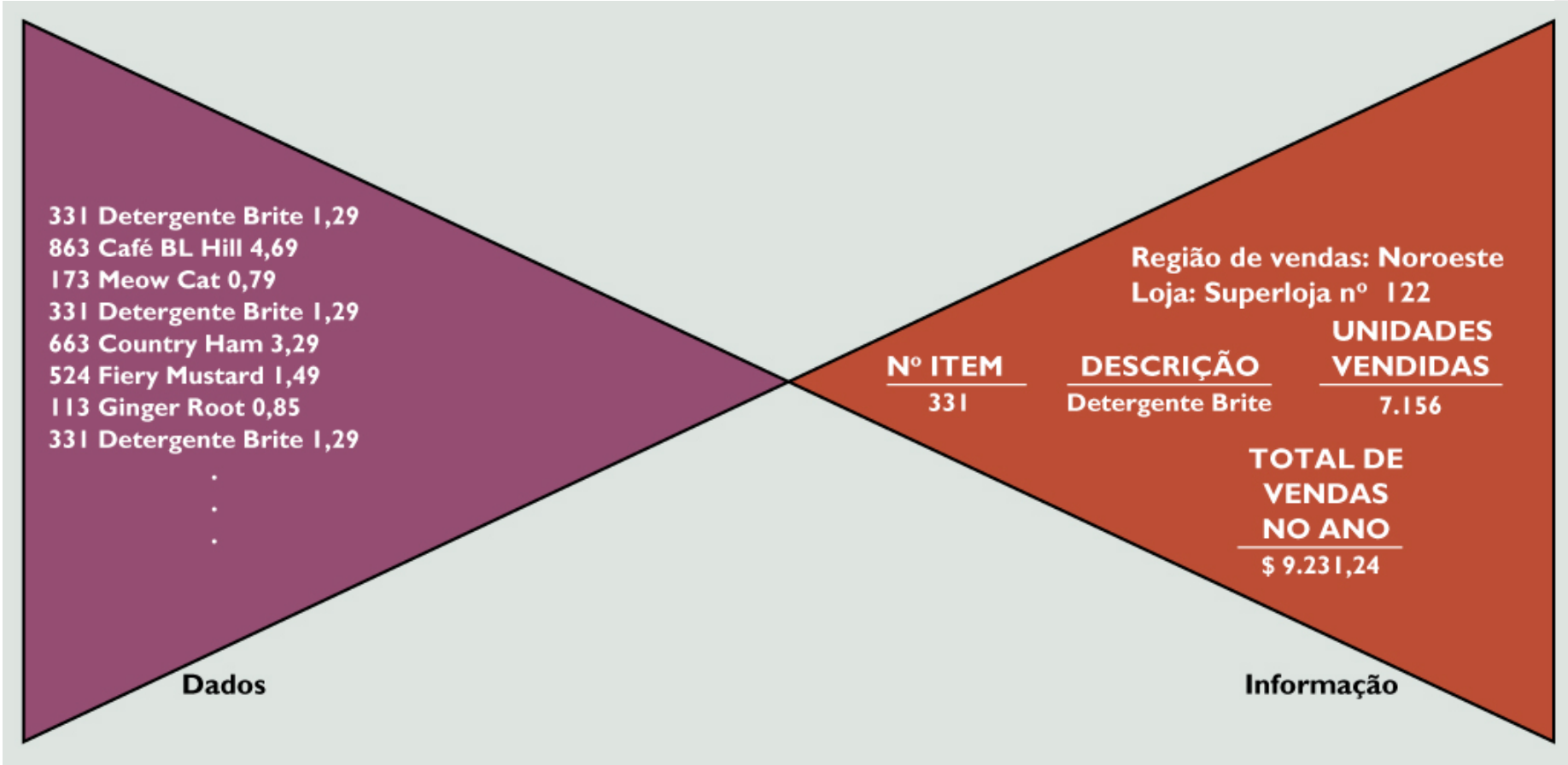
1. **Dados:** Fluxos de fatos coletados (brutos) que representam eventos do domínio. Qualquer evento que possa ser armazenado em formato digital, incluindo texto, números, imagens, vídeo ou filmes, áudio, software, algoritmos, equações, animações, modelos, simulações, etc.

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

O conceito de Dado x Informação x Conhecimento



Quais são os dados ?

Coleções de registros ou medições que fornecem um registro de evidências do evento observado *"... qualquer informação que possa ser armazenada em formato digital, incluindo texto, números, imagens, vídeo ou filmes, áudio, software, algoritmos, equações, animações, modelos, simulações, etc. "*

Atributos gerais da Iinformação:

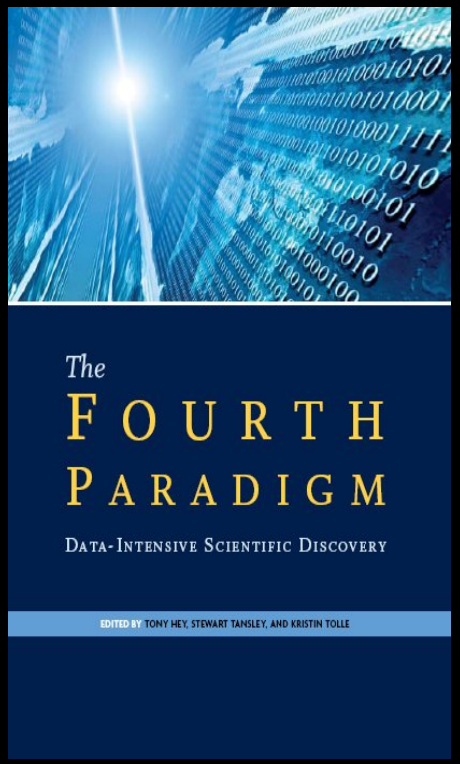
- Digital;
- Heterogêneo;
- Contextualizado;
- Valioso.

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

O conceito de Dado x Informação x Conhecimento

Pergunta:

Os dados climáticos coletados por estações metrológicas numa região é dado ou informação ?



BIG WORLD

BIG PROBLEMS

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.

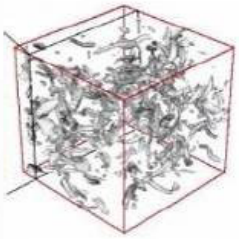
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$

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).

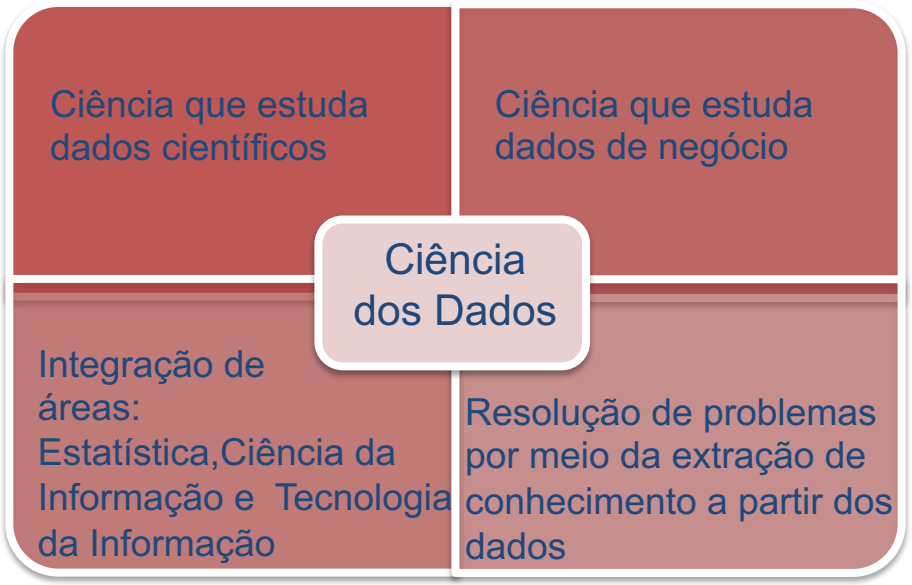
Três atividades consideradas na exploração de dados:
– Captura
– Curadoria
– Análise

- 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 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

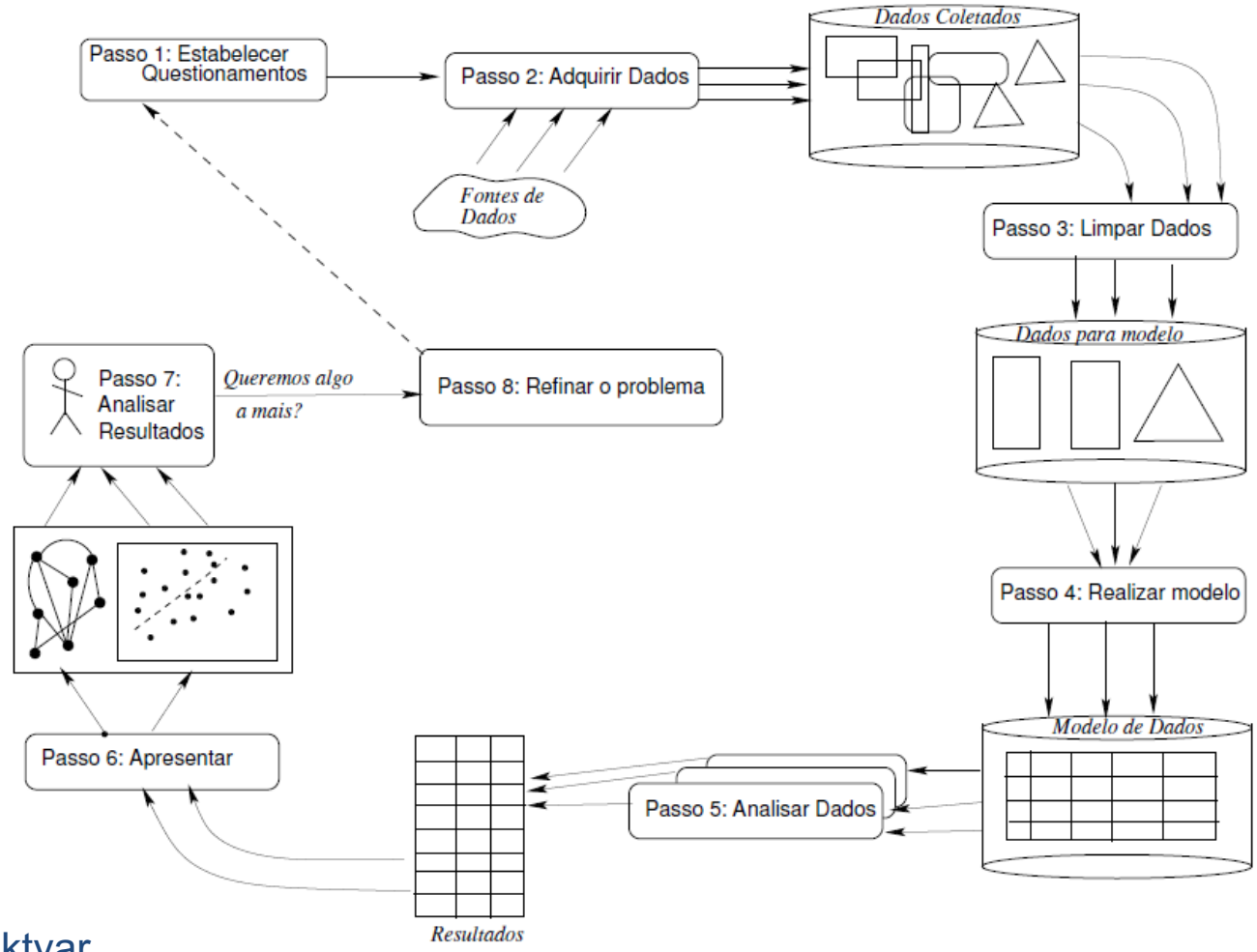


Ciência dos Dados

- 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

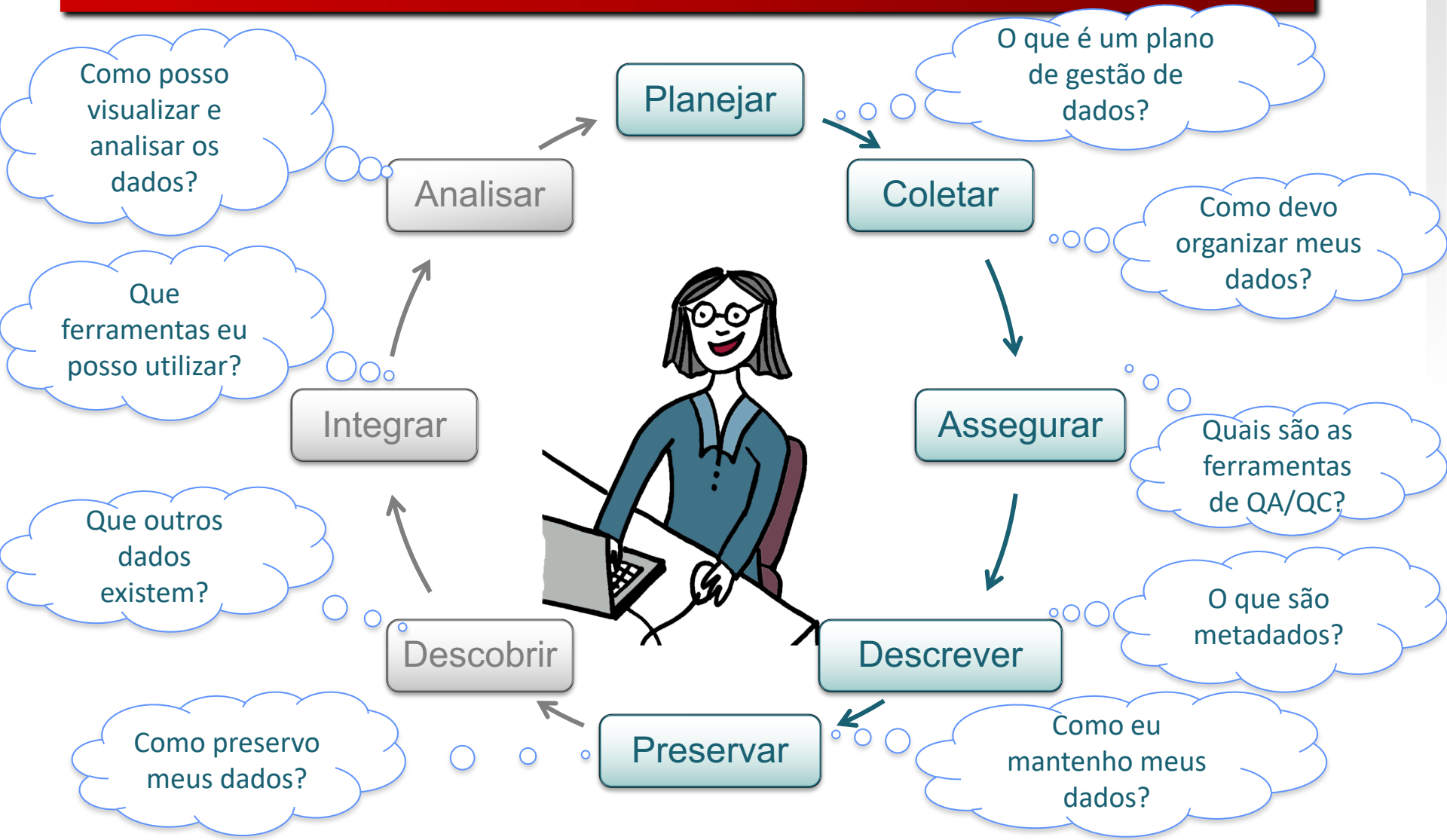
- Para Alex Dehktyar (2016),
  - Ciência dos dados é uma disciplina que permite tratar o o ciclo de trabalho com os dados, considerando atividades que compreendem desde a aquisição dos dados, passando pela análise dos dados, até o processo de apresentação dos dados e obtenção de novos conhecimentos

Ciência dos Dados - Processo



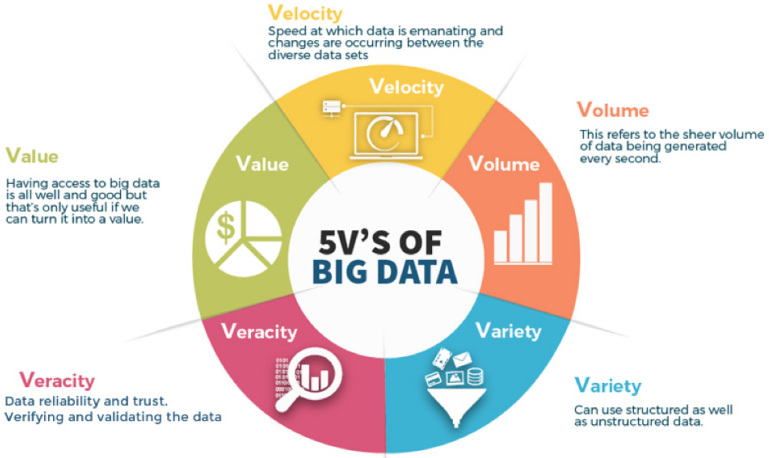
Cortesia: Alex Dehktyar

Ciência dos Dados - Gestão de Dados

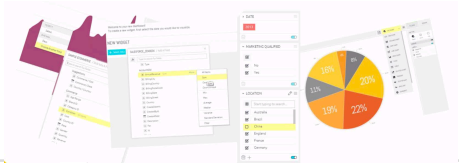


O que é Big Data ?

- ✓ É uma aplicação computacional de Ciência dos Dados que tem por objetivo analisar, extrair sistematicamente informações de grandes volumes de conjuntos de dados, para os quais técnicas computacionais tradicionais não são adequadas. Os desafios para gestão dos dados são classificados em 5V's (Chen et al., 2012, Kwon et al., 2014).



- ✓ Big data é um grande volume de dados, alta velocidade e alta variedade de ativos de informação que demandam formas inovadoras e econômicas de processamento de informações para melhor *insight* e tomada de decisões." ("Gartner IT Glossary, n.d.")



Causas que tornam os dados complexos

- Linguagem de consulta
- Tamanho
- Estrutura

- Dispersão
- Taxa de crescimento

Etapas para criação de Big Data

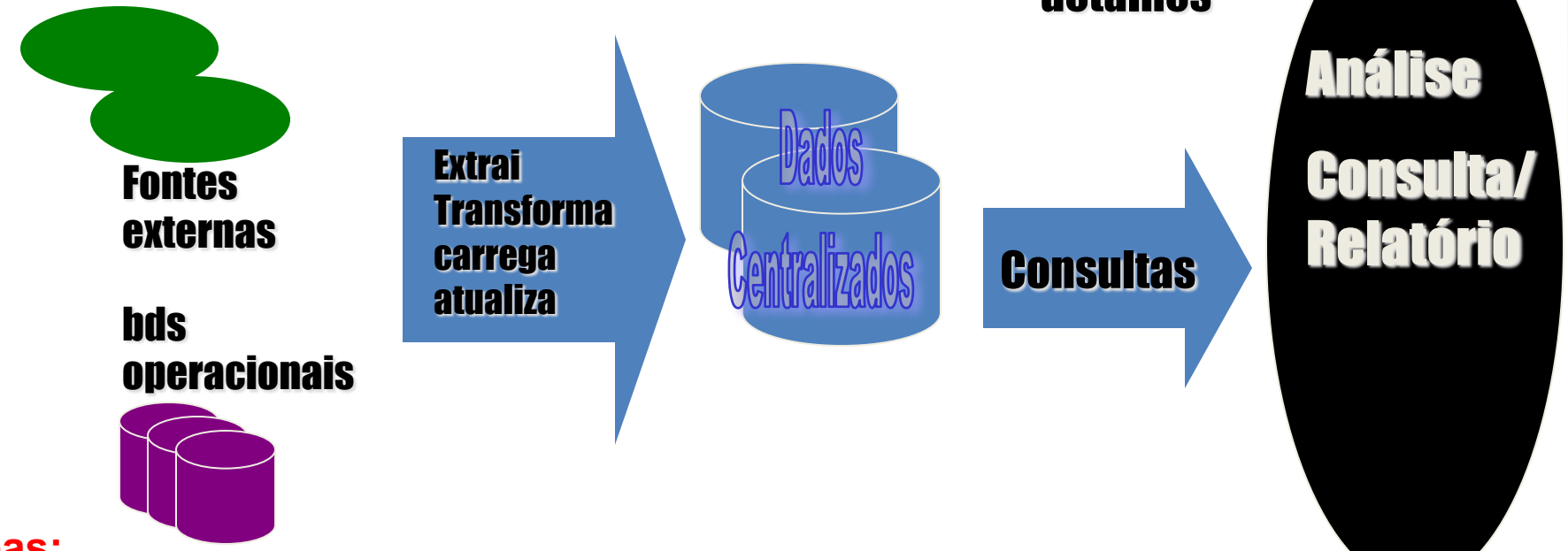
Questões a tratadas:

Dispersão

Estrutura

Tamanho e taxa de crescimento

Linguagem de consulta e detalhes



Etapas:

Fontes de Dados ETL

Centralização

Análise

Apoio a tomada de decisões

- Apoio para tomar decisão:
  - *Quais foram os volumes de vendas por região e categoria de produto no último ano ?*
  - *Quais pedidos devem ser priorizarados para maximizar os lucros ?*
  - *10% de desconto aumentará o volume de vendas suficientemente?*
- On-Line Analytical Processing (OLAP) é uma ferramenta de Sistemas de Suporte a Decisão (Decision Support Systems - DSS) e Sistema de Apoio a Executivos (SAE ou Executive Information System -EIS)

Evolução:

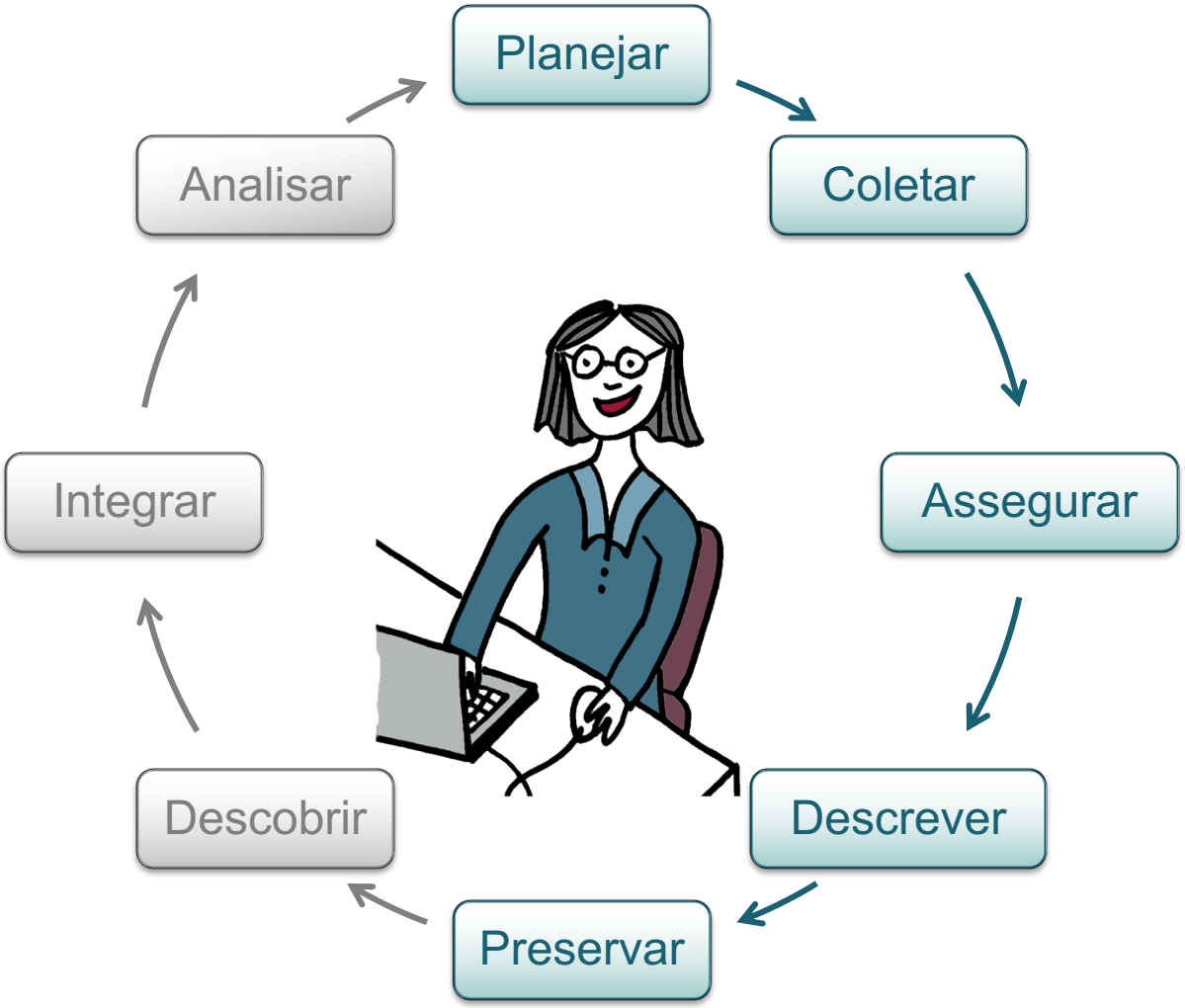
- 60: Relatórios Batchs
  - Longos relatórios;
  - Inflexíveis e caros, necessitando reprogramar a cada nova requisição
- 70: DSS e EIS usando terminais
  - Ainda Inflexível;
  - Não integrados a outras ferramentas de *desktop*.
- 80: Acesso através de *Desktop* e ferramentas de análise:
  - Ferramentas de consultas (*querys*), planilhas e GUI;
  - Fácil de usar, mas acessa somente a bases de dados operacionais
- 90: *Data warehousing* com OLAP e outras ferramentas integradas.
- 00/10: Aplicações/Serviços Web e Big Data

Agenda

- Introdução
  - O conceito de Dado x Informação x Conhecimento
  - Big Data, Ciência dos Dados, Gestão dos Dados
- Gestão dos dados – Importância e Desafios
- O modelo de dados do ponto de vista do *Big Data*
- Desafios relacionados ao *Big Data*
- Exercícios

Gestão de Dados – Importância e Desafios

Gestão de Dados



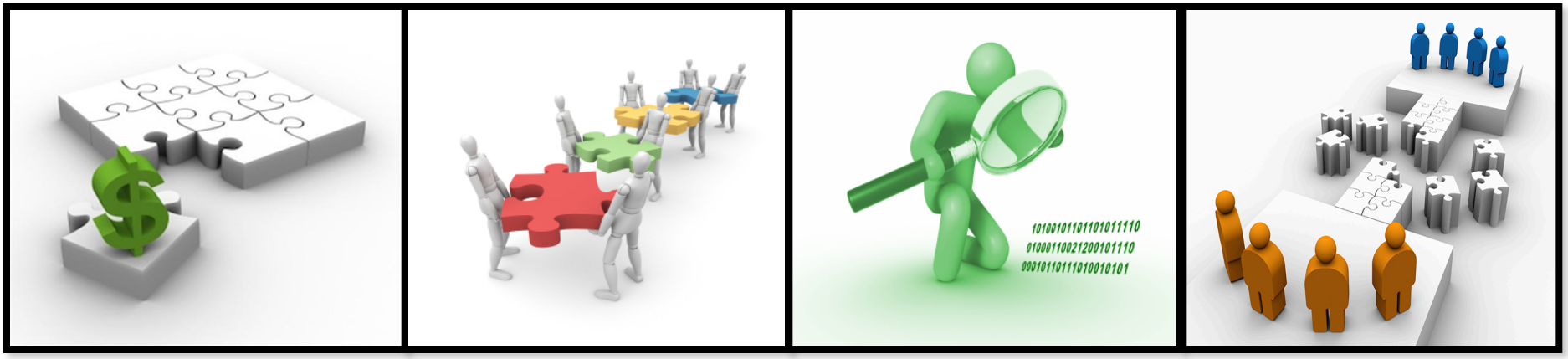
2. Por que a gestão de dados?

Porque aplicar as técnicas e conceitos de Gestão de Dados?

2. Porque a gestão de dados?

1. **Para capturar, armazenar, proteger e garantir a integridade dos ativos de dados;**
2. **Garantir a utilização adequada dos dados e informações;**
3. **Maximizar o uso eficaz dos dados e agregar valor aos ativos da informação.**

Fonte: DAMA International, *The DAMA Guide to the Data Management Body of Knowledge*



2. Porque a gestão de dados?

Se seus dados cairem em mãos erradas?

national security has leaked from Whitehall, the head of the civil service has warned.

Cabinet Secretary Sir Gus O'Donnell said there were "one or two" leaks from areas dealing with national security.

However he ... sensitive official ... have hit the headlines

... been a series of breaches in recent years, including of the entire child benefit records, with the personal details of 25 million people.

Giving evidence to the Commons Public Administration Committee, Sir

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?

change row

Leading British scientists at the University of East Anglia, who were accused of manipulating climate change data - dubbed Climategate - have agreed to publish their figures in full.

Print this

Twitter 3

Email

LinkedIn 0

Copenhagen climate change conference

News » UK News » Earth News »

CIENTISTAS FORAM ACUSADOS DE MANIPULAR DADOS SOBRE MUDANÇAS CLIMÁTICAS.

Fonte: The Telegraph

2. Por que a gestão de dados?

Se este for o seu Instituto de Pesquisa?

Incêndio no Instituto Butantan destrói maior acervo de cobras do país

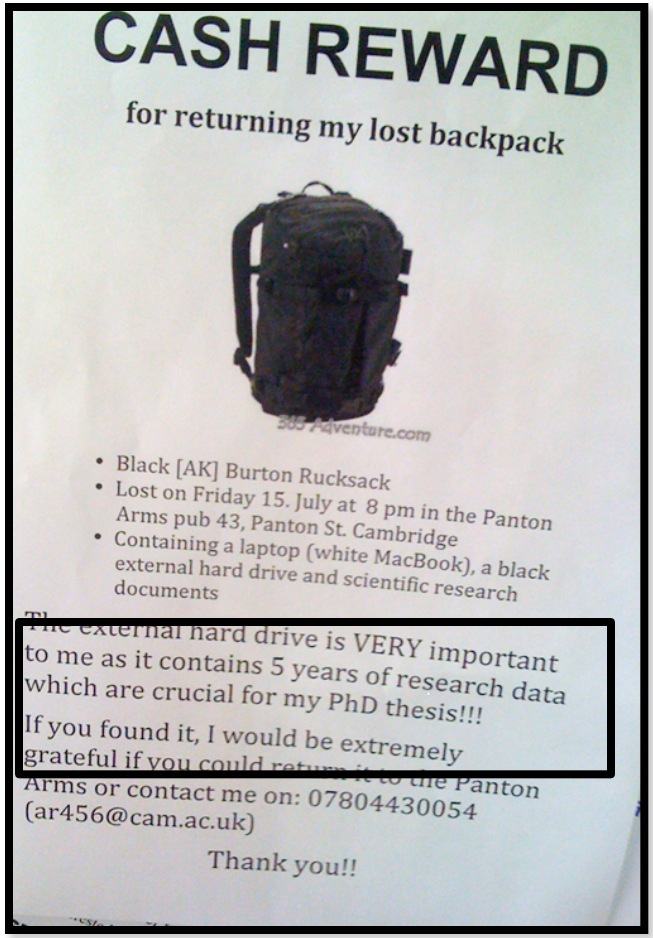
Fogo queimou 70 mil espécies conservadas em formol na Zona Oeste de SP. Chamas atingiram laboratório de répteis; causas do fogo serão apuradas.

Fonte: http://g1.globo.com/sao-paulo/noticia/2010/05/incendio-no-instituto-butantan-destroi-maior-acervo-de-cobras-do-pais.html

2. Por que a gestão de dados?

SE ESTA FOR A SUA MOCHILA?

"O HD externo é muito importante, pois contém 5 anos de backup de e-mails"



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

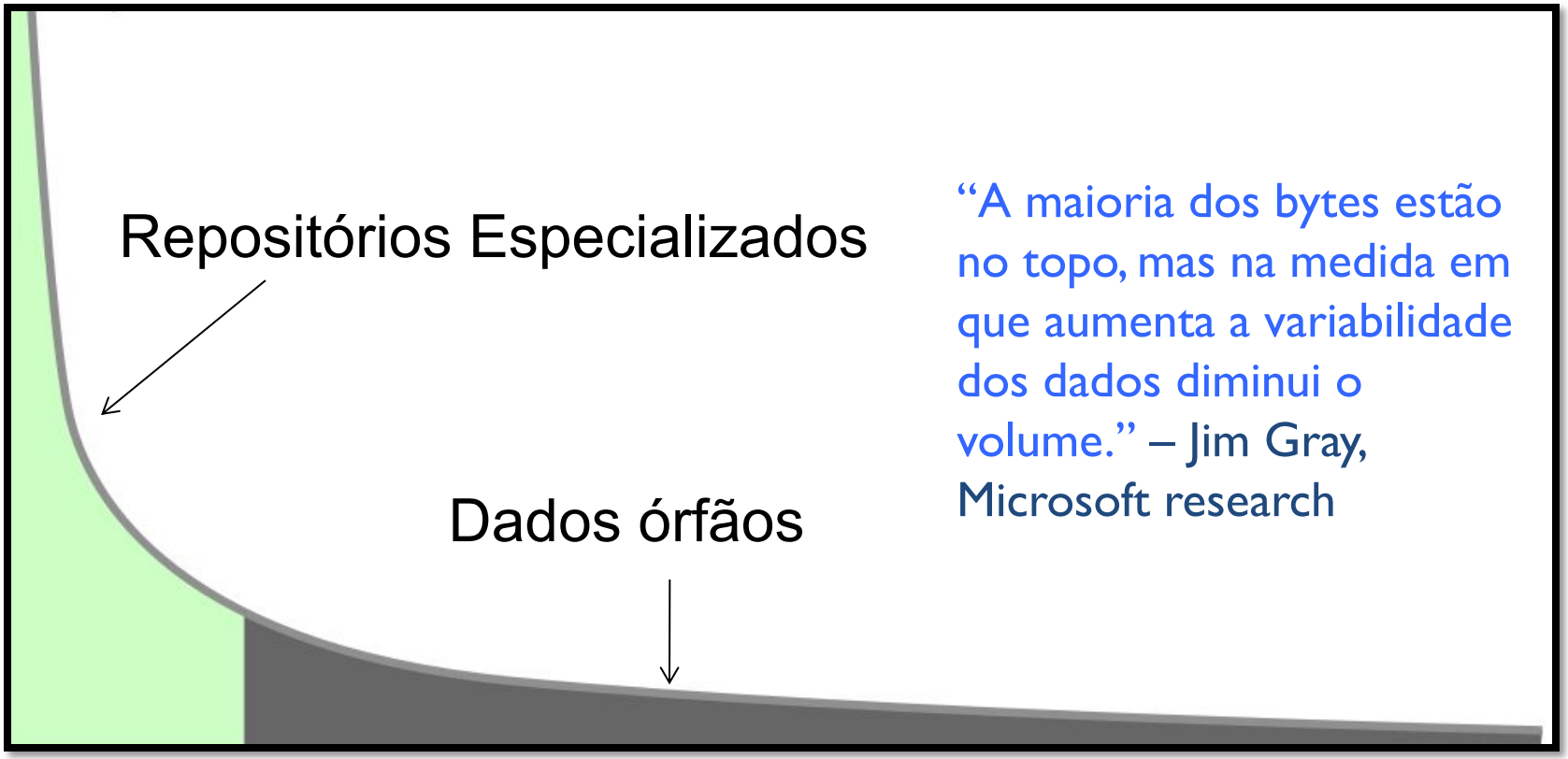
2. Gestão de dados

“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: "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

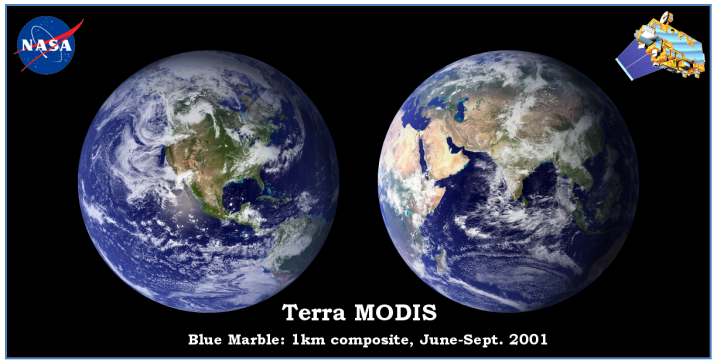
Fonte: DataONE

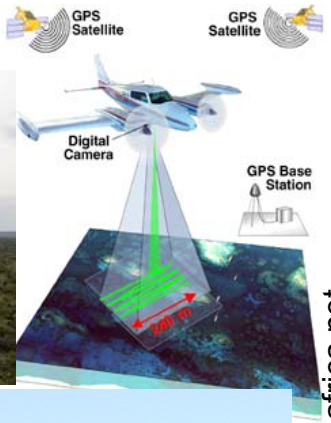
2. Desafios: Dados órfãos

- 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 de dados.

2.Desafios: "Dilúvio" dos Dados

Redes, Sensores, Sensoriamento Remoto, Experimentos, Coletas…



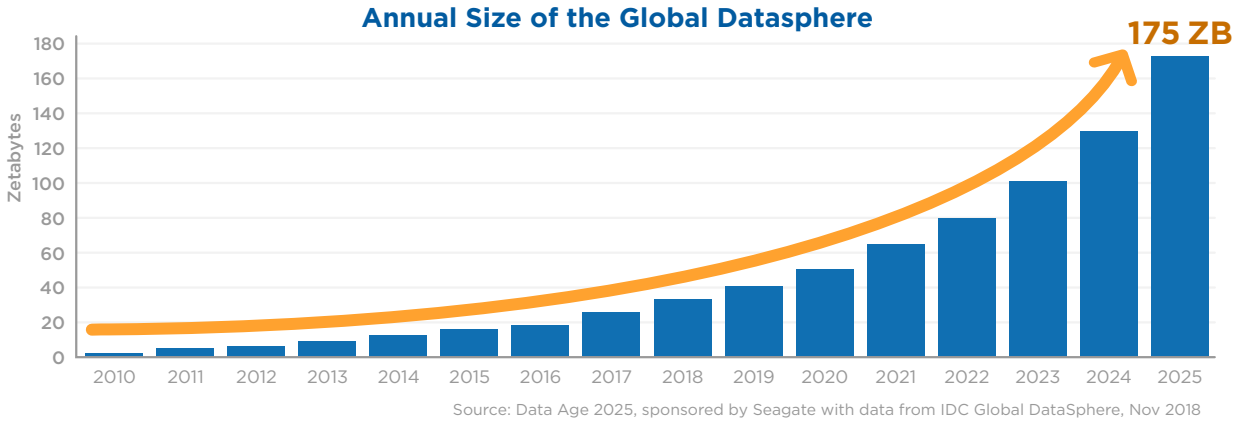


Fonte: www.carboafrica.net

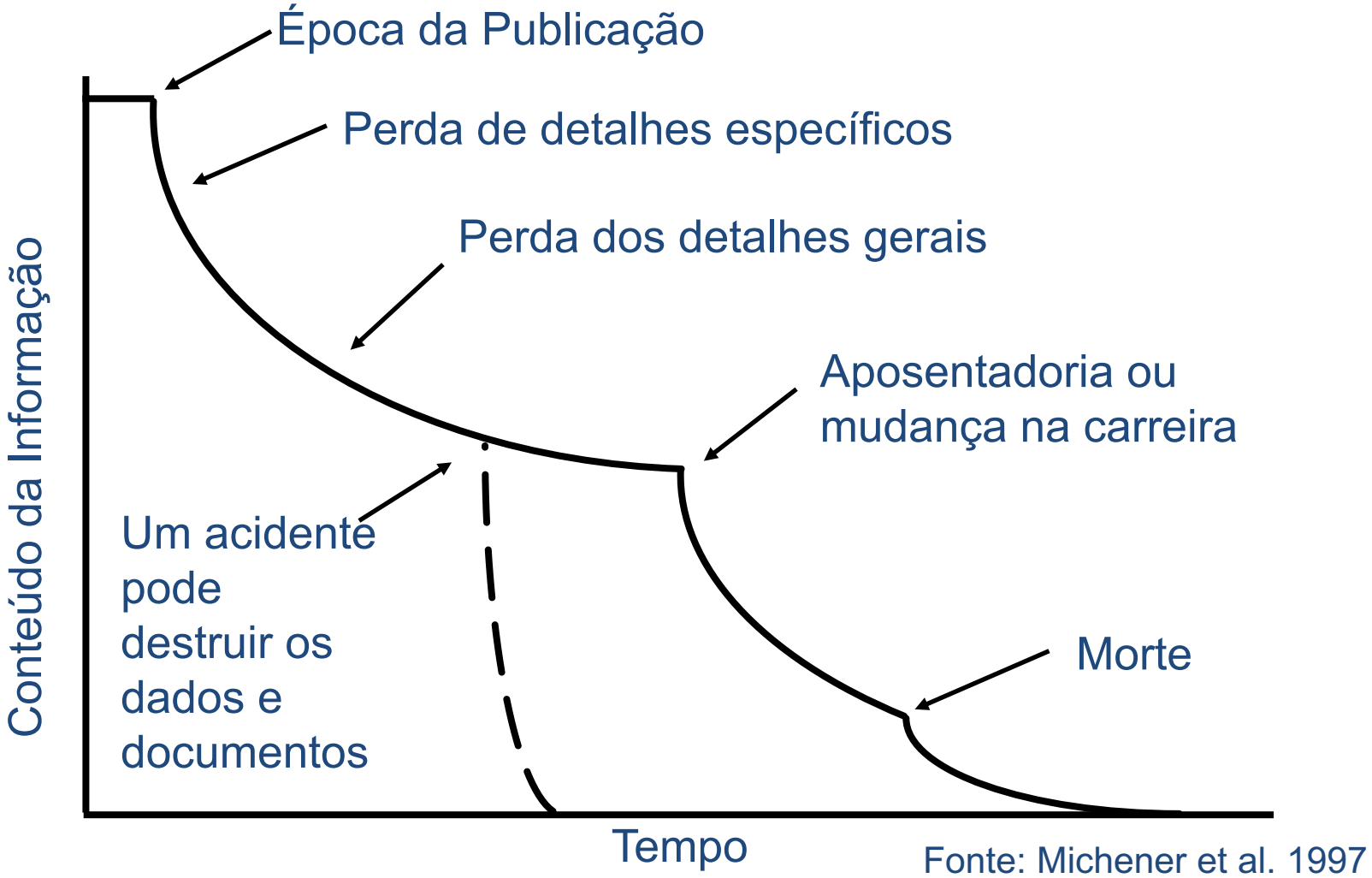
Problema: Volume de Dados

International Data Corporation (IDC):

- A informação no mundo está mais que dobrando, está dobrando cada dois anos (30 Zettabytes em 2018);
- O número de arquivos está crescendo mais rápido que a capacidade de armazenamento. Nos próximos 5 anos esses arquivos irão crescer num fator de 8x;
- O número de pessoas na área de IT responsável por esses dados irá crescer "suavemente";

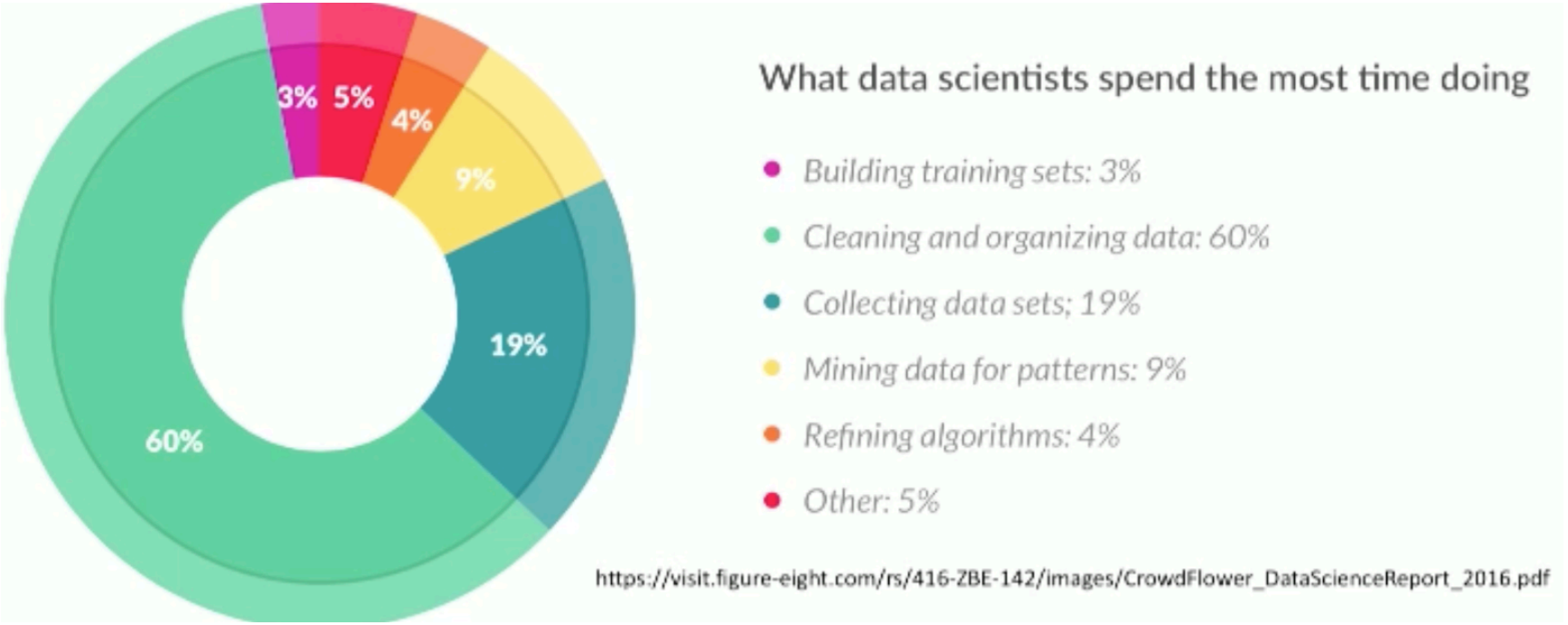


2.Desafios: Entropia dos Dados



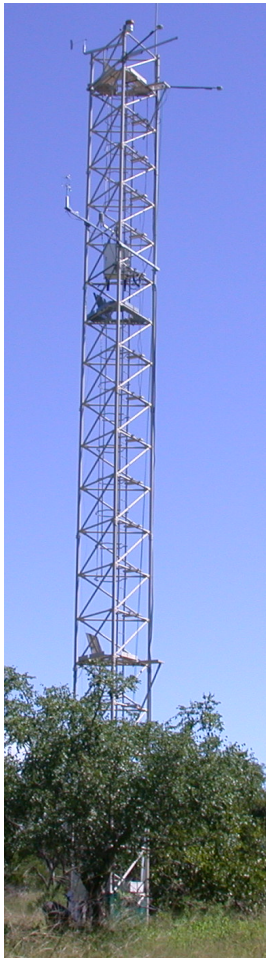
Fonte: Michener et al. 1997

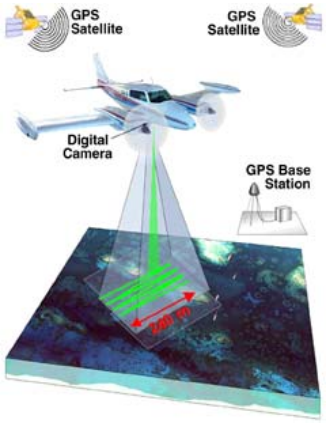
Problema que conhecemos muito bem !

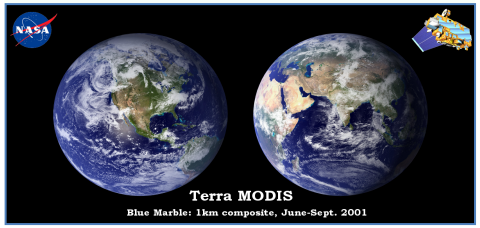


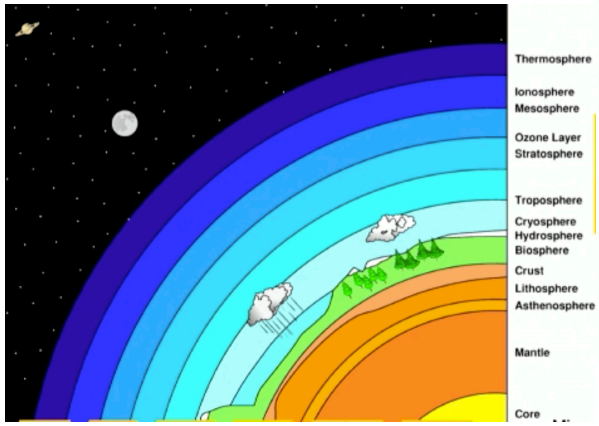
- 60% em atividades na organização e limpeza dos dados;
- 19% em atividades de pré-processamento
- 9% em atividades de análise

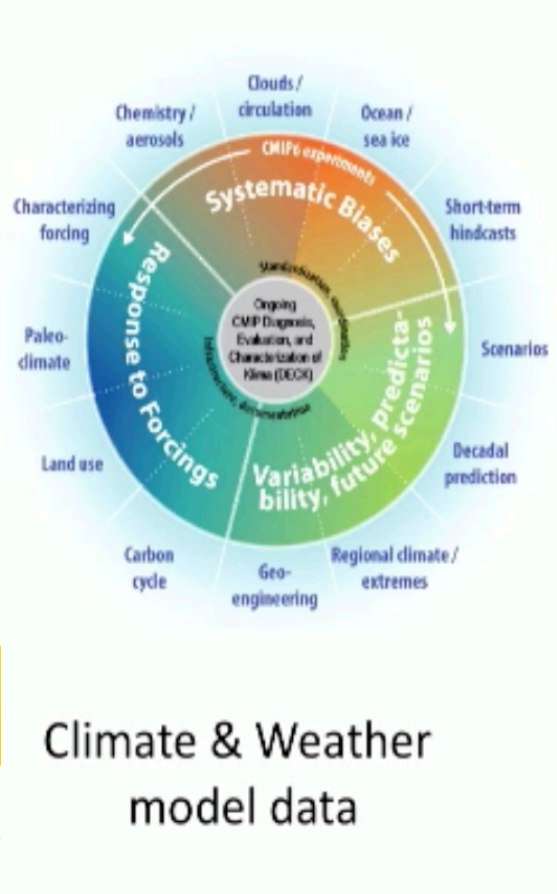
Quais são os vetores desse dilúvio de dados (data deluge)











Agenda

- Introdução
  - O conceito de Dado x Informação x Conhecimento
  - Big Data, Ciência dos Dados, Gestão dos Dados
- Gestão dos dados – Importância e Desafios
- O modelo de dados do ponto de vista do *Big Data*
- Desafios relacionados ao *Big Data*
- Exercícios

Gestão de Dados – Boas Práticas e Modelo

2. Boas Práticas na Gestão de Dados

Planejar → **Coletar**

Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

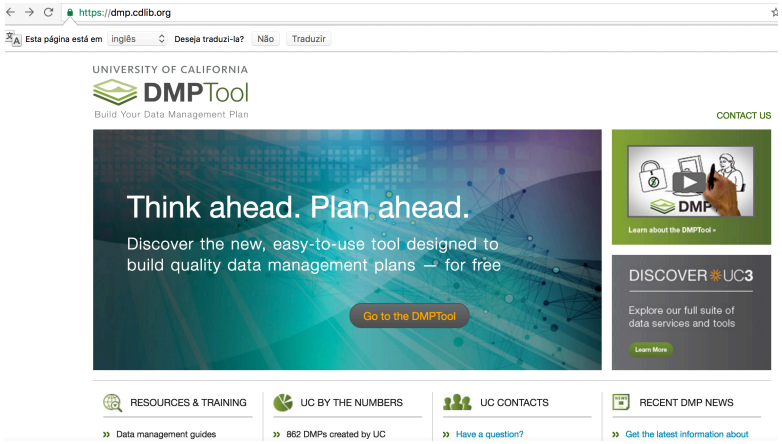
2 - Planejar

a) Mapeia os processos e recursos para todo o ciclo de vida dos dados;

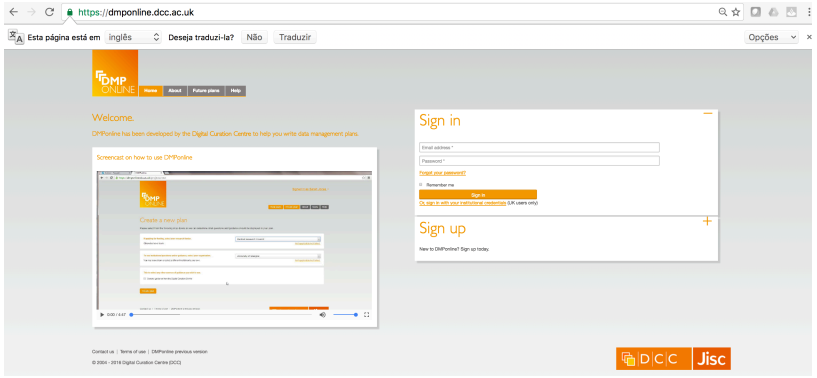
- Começar com os objetivos do projeto (desejos, resultados e impactos) e construir um plano de gestão dos dados, considerando a política dos dados e sua sustentabilidade.
- Pontos a serem considerados:
  - Quais e como os dados serão Coletados;
  - Onde os dados serão armazenados (Repositório);
  - Como os dados serão organizados (formatos, estrutura);
  - Como os dados serão descritos (metadados);
  - Como os dados serão compartilhados;
  - Definição do plano de preservação dos dados e responsabilidades.

2 - Plano de Gestão dos Dados

b) Ferramentas para a criação de Planos de Gestão de Dados:

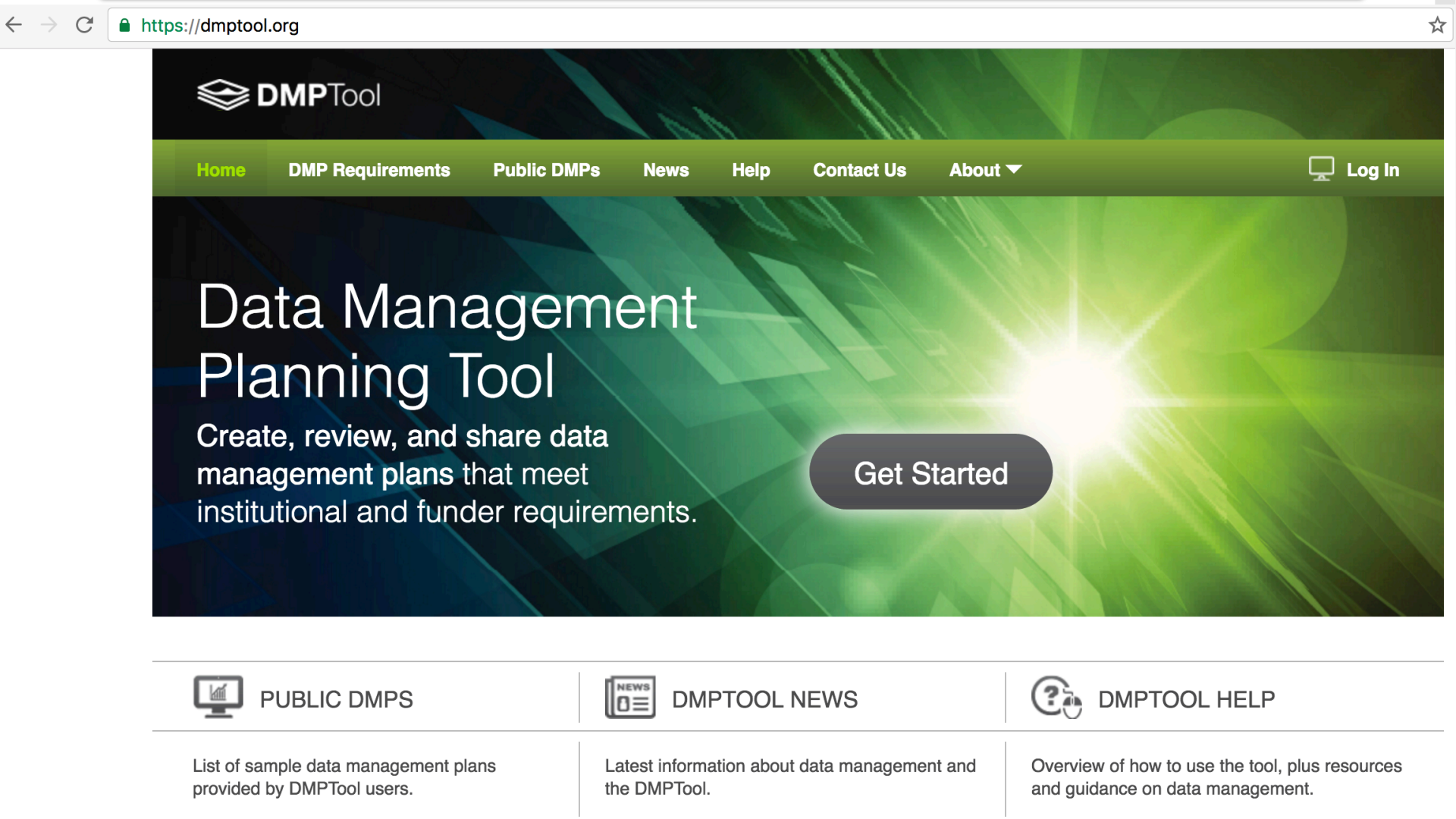


dmp.cdlib.org



dmponline.dcc.ac.uk

2 - Plano de Gestão dos Dados



Boas Práticas na Gestão de Dados

Planejar → Coletar → Assegura

Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

2 - Coletar

1. Determina a melhor forma para se obter os dados;
2. O resultado deste processo é um Documento que descreve a forma como os dados são estruturados.
   - Determina como os dados são coletados;
   - Determina como os dados gerados são organizados e armazenados.
3. Para o armazenamento em longo prazo, os dados devem ser gerados em formatos de dados consistentes e específicos para esta finalidade (facilitando sua utilização hoje e no futuro).
   - Bases Relacionais;
   - Arquivos CSV (Texto com separadores);
   - Outros recomendações de formatos
     https://libraries.mit.edu/data-management/store/formats/

Problemas na coleta

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	riuSlado	"	1*	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Ccmments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

- Inconsistências entre os eventos coletados;
- Localização das informações de datas;
- Inconsistência no formato de datas;
- Nome das colunas;
- Ordem das colunas.

Problemas na coleta

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping	3/15/2010					
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	rioSlado	"	1*	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Comments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

Sheet1

SEV_SmallMammalData_v.5.25.2010.xls

	A	B	C	D	E	F	G	H
1	Date	Site	Plot	Species	Weight	Adult	Comments	
2	2/5/2010	Deep Well	1	DIPO	13.2	y		
3	2/4/2010	Deep Well	1	CLEGAP	11.6	j		
4	2/5/2010	Rio Salado	1	DIPO	14.2	y		
5	2/5/2010	Rio Salado	2	PERO	10.1	y		
6	3/15/2010	Deep Well	1	DIPO	15.2	y	plot burned	
7	3/15/2010	Deep Well	2	DIPO	21.7	y	pregnant	
8	3/15/2010	Rio Salado	1	CLEGAP	16.2	j		
9								
10								
11								
12								

SmallMammalTrapping | Sheet3

- Colunas devem ser consistentes: apenas números, datas ou textos;
- Formatos e códigos consistentes;
- Dados em apenas uma planilha, para facilitar a captura dos dados por sistemas computacionais, sem a intervenção humana;
- Colunas específicas para comentários e demais características que descrevem o registro.

Boas Práticas na Gestão de Dados

Coletar → **Assegurar** → Descrever

Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

3 - Assegurar

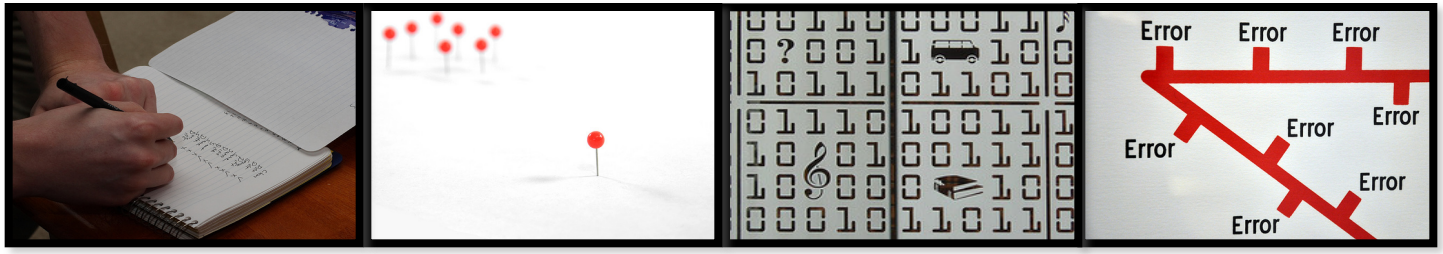
1. Empregar procedimentos de garantia e controle da qualidade (QA/QC) dos dados:
   - Capacitação de Coletores;
   - Rotina de calibração de instrumentos;
   - Procedimentos de revisão dos dados.
2. Identificar problemas e técnicas computacionais possíveis para solucioná-los.

Quality Assurance (QA) é um conjunto de processos utilizados para garantir que os dados necessários serão coletados e armazenados;

Quality Control (QC) é um conjunto de processos para avaliar a qualidade dos dados depois que os mesmos forem coletados.

3-Assegurar

1. **Tipos mais comuns de erros:**
   - Erros de Ação
     - Dados inseridos incorretamente;
     - Exemplos: Instrumentos mal calibrados, erros de digitação.
   - Erros de Omissão
     - Dados ou Metadados não registrados;
     - Exemplos: Documentação inadequada, erro humano, anomalias nas coletas.



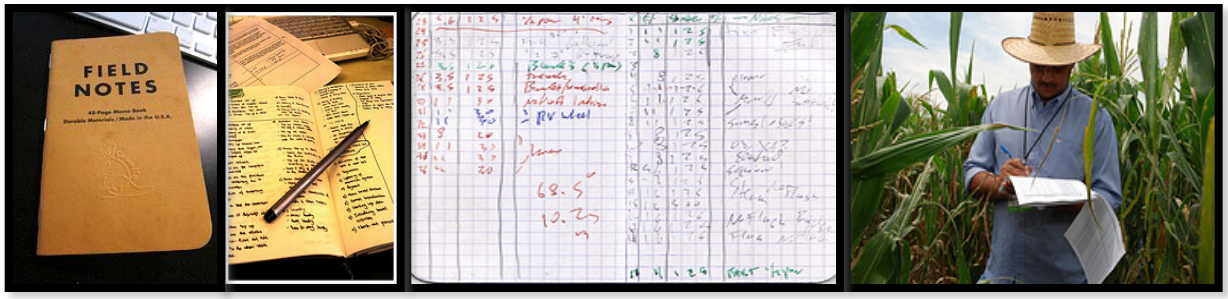
Boas Práticas na Gestão de Dados

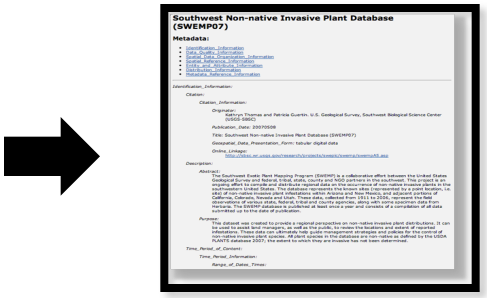
Assegurar → **Descrever** → Preservar

Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

4 - Descrever

1. **Documentação dos dados**
   - Porque, quem, o quê, quando, onde e como;
   - Utilização de Metadados (dados sobre dados);
     - É a chave para o compartilhamento e reutilização dos dados.
     - Várias normas e ferramentas estão disponíveis para apoiar este processo.
     - Os metadados são utilizados tanto por humanos quanto por computadores para apoio nos processos de descoberta, integração e análise dos dados.





4- Descrever

1. **Reutilização dos dados**
   - **Se são criados metadados, outras pessoas podem descobrir seus dados.**
   - **Se são criados metadados, você pode encontrar os seus próprios dados!**

Boas Práticas na Gestão de Dados

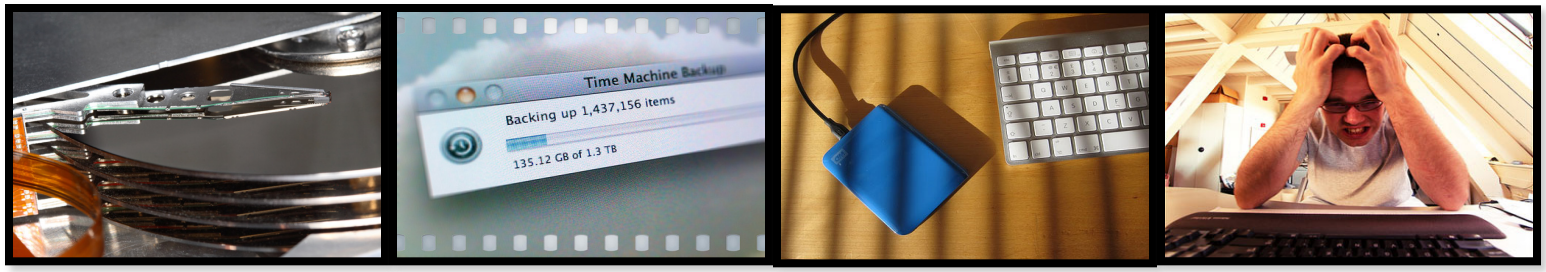
Descrever → **Preservar** → Descobrir

Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

5 - Preservar

Plano para preservar os dados

- No curto prazo para minimizar as perdas potenciais, como acidentes;
- No Longo prazo para que participantes do projeto e outros pesquisadores possam acessar, interpretar e utilizar os dados no futuro.
- Proteção dos dados incluem questões como backups, segurança física, criptografia;
- Preservação dos dados incluem os processos para conservação, recuperação, reorganização e descrição dos dados.



5 – Preservar

Boas práticas

- **Armazenar os dados e metadados em formatos apropriados;**
- **Realizar backups dos dados dos projetos;**
- **Garantir a integridade e o acesso aos backups de dados;**
- **Definição de políticas para segurança e armazenamento das informações;**
- **Identificar dados com valor a longo prazo;**
- **Preservar os dados brutos;**
- **Identificar e gerenciar os dados sensíveis.**

5 – Preservar - exemplo

- DDOMP Checklist – Team Resources

- Material development and temporary storage location
  - Google Drive
- Team communications and information decimation tools
  - Email, Slack
- Dataset storage location during the project
  - Open Science Framework (AWS integration)
- Software development platform
  - GitHub
- Data preservation (including derived products) repository
  - Environmental Data Initiative
- Software preservation repository
  - Zenodo
- Training, workshop material preservation repository
  - Zenodo

PARSEC

- PIs – 4
- Country Leaders – 6
- Funders – 4
- Researchers – 30
- Languages - 4

- Stall, S. DDOMP - 10.5281/zenodo.3942688
- Projeto Parsec: https://parsecproject.org/

5 – Preservar - exemplo

- DDOMP Checklist – Tracking and Reporting

- **Once (during lifetime of researcher)**

- **ORCID profile** - activate the automatic updates from Crossref (published papers) and DataCite (published datasets and other digital objects. Reference this page for instructions: http://bit.ly/ORCID_Trust

- **Weekly**

- Track **datasets created**, track **datasets used**, track **workflow/provenance**

- **Monthly**

- Publish and Report **conference presentations and posters**
- Peer-reviewed Papers – and Supporting Digital Objects
  - Deposit and preserve **Datasets**
  - Deposit and preserve **Software**
  - Report **publications with citations to datasets, software, and other digital objects**

- **Quarterly**

- - Update your ORCID profile and ensure accurate and complete to ensure proper credit

- Stall, S. DDOMP - 10.5281/zenodo.3942688
- Projeto Parsec: https://parsecproject.org/

5 – Preservar - exemplo

- PARSEC Data and Digital Output Management Plan and Workbook

Further details can be found on the process and methods used for PARSEC:

Stall, Shelley, Specht, Alison, Corrêa, Pedro Luiz Pizzigatti, David, Romain, Edmunds, Rorie, Mabile, Laurence, Machicao, Jeaneth, O'Brien, Margaret, Wyborn, Lesley. (2020). PARSEC Data and Digital Output Management Plan and Workbook. Zenodo.

10.5281/zenodo.3891426

Use your DMP or DDOMP to make your own Checklist.

Boas Práticas na Gestão de Dados

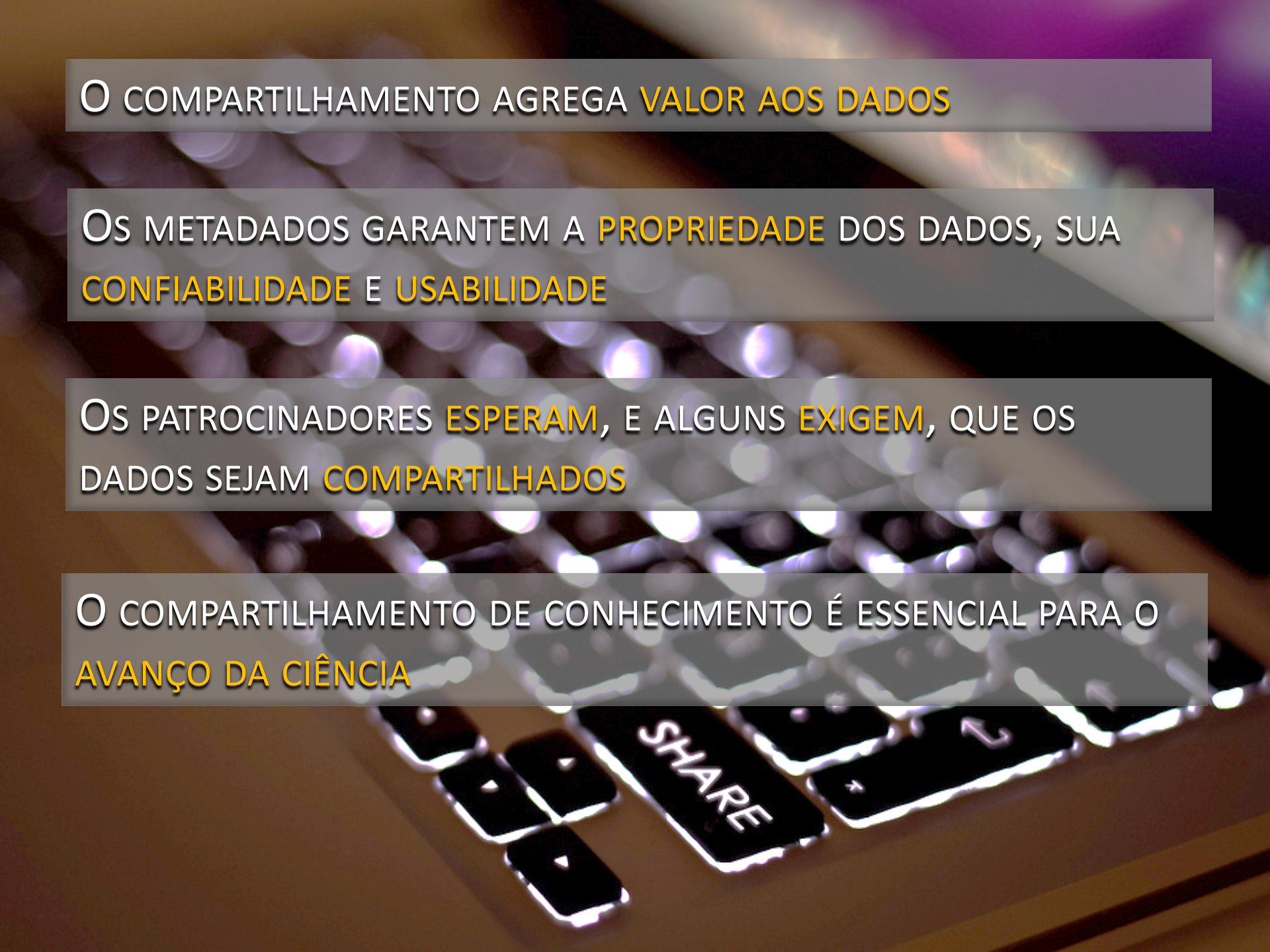
Preservar → **Descobrir** → Integrar

Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

6 - Descobrir

1. **Estratégias para a localização e aquisição de dados potencialmente úteis:**
   - **Identificar dados complementares que possam agregar valor aos dados do projeto.**
   - **Buscadores genéricos na Web não são eficientes para encontrar dados úteis.**
   - **Dados são mais facilmente encontrados por meio das redes, diretórios de projetos e repositórios:**

Repositórios	Website
Global Biodiversity Information Facility	gbif.org
Atlas of Living Australia	ala.org.au
Knowledge Network for Biocomplexity	knb.ecoinformatics.org
Dryad	datadryad.org
DataONE	search.dataone.org



O COMPARTILHAMENTO AGREGA VALOR AOS DADOS

OS METADADOS GARANTEM A PROPRIEDADE DOS DADOS, SUA CONFIABILIDADE E USABILIDADE

OS PATROCINADORES ESPERAM, E ALGUNS EXIGEM, QUE OS DADOS SEJAM COMPARTILHADOS

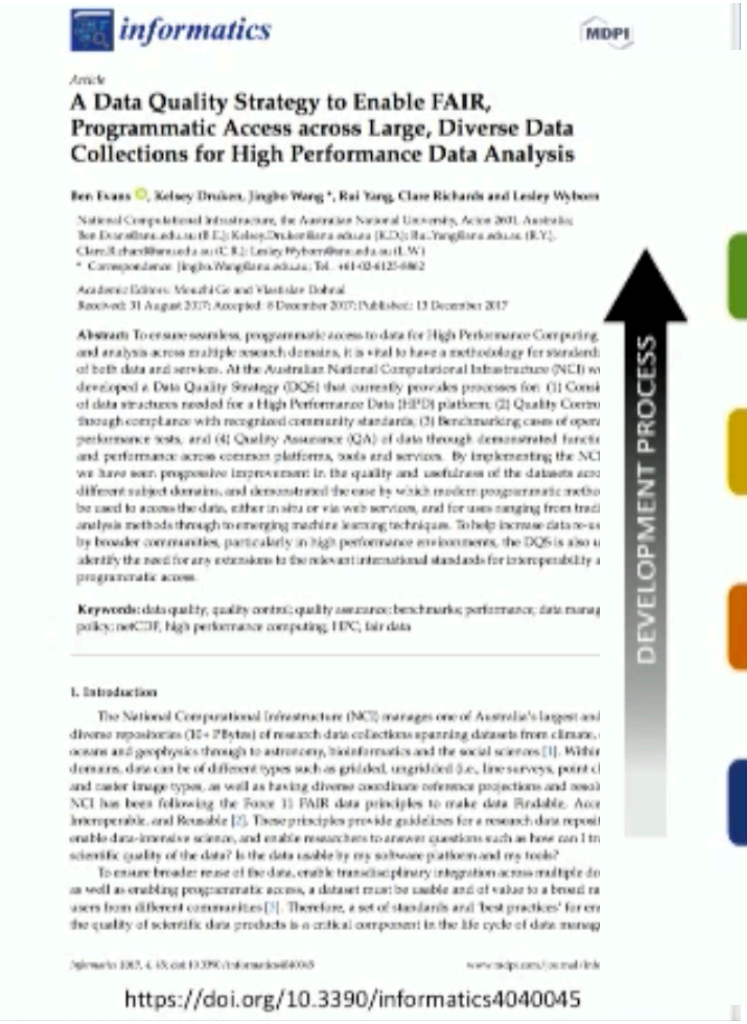
O COMPARTILHAMENTO DE CONHECIMENTO É ESSENCIAL PARA O AVANÇO DA CIÊNCIA

Boas Práticas na Gestão de Dados

Descobrir → **Integrar** → **Analisar**

Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

FAIR Data Principles
to be Findable, Accessible, Interoperable and Re-Usable



informatics

MDPI

Article

A Data Quality Strategy to Enable FAIR, Programmatic Access across Large, Diverse Data Collections for High Performance Data Analysis

Ben Evans, Kelsey Druken, Jingbo Wang *, Rui Yang, Clare Richards and Lesley Wyborn

National Computational Infrastructure, the Australian National University, Acton 2601, Australia; Ben.Evans@anu.edu.au (B.E.); Kelsey.Druken@anu.edu.au (K.D.); Rui.Yang@anu.edu.au (R.Y.); Clare.Richards@anu.edu.au (C.R.); Lesley.Wyborn@anu.edu.au (L.W.)
* Correspondence: Jingbo.Wang@anu.edu.au; Tel.: +61-02-6125-6862

Academic Editors: Wenzhi Cao and Vladislav Bohrial
Received: 31 August 2017; Accepted: 8 December 2017; Published: 13 December 2017

Abstract: To ensure seamless, programmatic access to data for High Performance Computing and analysis across multiple research domains, it is vital to have a methodology for standardized of both data and services. At the Australian National Computational Infrastructure (NCI) we developed a Data Quality Strategy (DQS) that currently provides processes for: (1) Consistency of data structures needed for a High Performance Data (HPD) platform; (2) Quality Controls through compliance with recognized community standards; (3) Benchmarking cases of operations performance tests; and (4) Quality Assurance (QA) of data through demonstrated functionality and performance across common platforms, tools and services. By implementing the NCI DQS we have seen progressive improvement in the quality and usefulness of the datasets across different subject domains, and demonstrated the ease by which modern programmatic methods can be used to access the data, either in situ or via web services, and for uses ranging from traditional analysis methods through to emerging machine learning techniques. To help increase data re-use by broader communities, particularly in high performance environments, the DQS is also used to identify the need for any extensions to the relevant international standards for interoperability and programmatic access.

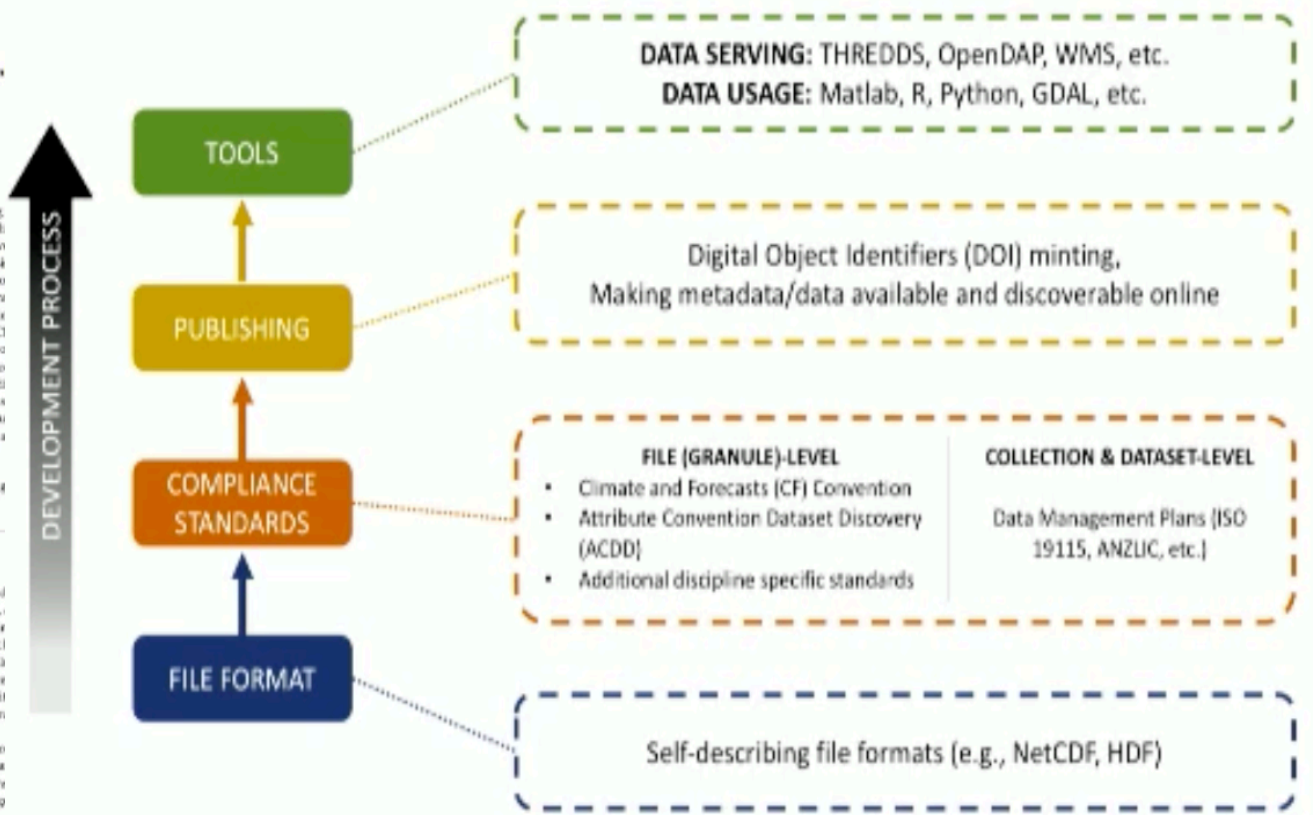
Keywords: data quality; quality control; quality assurance; benchmarks; performance; data management policy; netCDF; high performance computing; HPC; fair data

1. Introduction

The National Computational Infrastructure (NCI) manages one of Australia's largest and most diverse repositories (10+ PBytes) of research data collections spanning datasets from climate, ocean and geophysics through to astronomy, bioinformatics and the social sciences [1]. Within these domains, data can be of different types such as gridded, ungridded (i.e., line surveys, point clouds) and raster image types, as well as having diverse coordinate reference projections and resolutions. NCI has been following the Force 11 FAIR data principles to make data Findable, Accessible, Interoperable, and Reusable [2]. These principles provide guidelines for a research data repository to enable data-intensive science, and enable researchers to answer questions such as how can I trust the scientific quality of the data? Is the data usable by my software platform and my tools?

To ensure broader reuse of the data, enable transdisciplinary integration across multiple domains as well as enabling programmatic access, a dataset must be usable and of value to a broad range of users from different communities [3]. Therefore, a set of standards and 'best practices' for ensuring the quality of scientific data products is a critical component in the life cycle of data management

https://doi.org/10.3390/informatics4040045



7 – Integrar/Publicar

Combina dados de diversas fontes para possibilitar novas análises e investigações.

- O sucesso da integração de dados depende do emprego de boas práticas de gestão em todo o Processo de Gestão dos dados.
- Existem diversos cenários para a integração de dados:
  - Integração de dados de múltiplos projetos para o tratamento de questões complexas;
  - Dados esparsos que precisam ser complementados com dados existentes para possibilitar a realização de análises;

Exemplos:

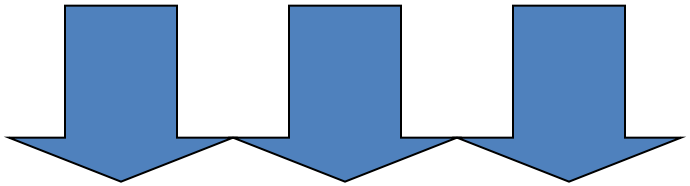
CKAN (Dados Abertos de Governo – http://ckan.org) , Quandl (Dados Financeiros - https://www.quandl.com), Kaggle datasets. (https://www.kaggle.com/datasets), Microsoft Azure Market Place (http://datamarket.azure.com/browse/data)

7-Integrar/Publicar

1. Estratégias para tornar as informações sobre os dados disponíveis, para que os outros possam descobri-los e acessá-los:
    1. Aumentando a visibilidade do projeto e de seus dados:
        - Aumento do potencial de uso amplo;
        - benefícios para a pesquisa científica, o apoio à decisão e a elaboração de políticas públicas.
    2. Estratégias para garantir o máximo impacto para os dados gerados:
        - Registrar o projeto em um site de diretório de projetos;
        - Depositar os dados gerados em um repositórios compartilhados;
        - Adicionar descrições dos dados (metadados) em sistemas de armazenamento distribuído de metadados.

7 - Esforço necessário

- Normalmente, essas atividades envolvem acesso a conjunto de dados independentes, assim:
  - **Usuários devem executar muitas atividades de gerenciamento de dados para extrair, integrar e analisar dados.**
- Necessário uma visão integrada através de todas as redes de dados, permitindo uma busca integrada, análise e visualização através de um conjunto comum de ferramentas e protocolos.



Abordagem de gerenciamento integrado do Ciclo de Vida dos Dados

7 – D.O.I. para Dados

Conceito

- DOI é um acrônimo para "*Digital Object Identifier*", o que significa um "identificador digital de um objeto". Um DOI é um identificador (não um endereço) de uma entidade em redes digitais.

- Um DOI pode ser atribuído a qualquer entidade (objeto) - físico, digital ou abstrato - principalmente para compartilhar em uma comunidade de utilizadores interessados ou na gestão como propriedade intelectual.

- DOI é expresso em URLs (URIs).

- DataCite: organização responsável por implementar D.O.I para dados.

7 – D.O.I. para Dados

Sintaxe

- A sintaxe do DOI especifica a construção de uma cadeia com autoridade e delegação de nomes. Ele fornece um "container" identificador que pode acomodar qualquer identificador existente.

- O DOI tem dois componentes, o prefixo e o sufixo, que juntos formam o DOI, separado pelo caractere "/".

- Não há qualquer limitação sobre o comprimento de um DOI.
- Exemplos:

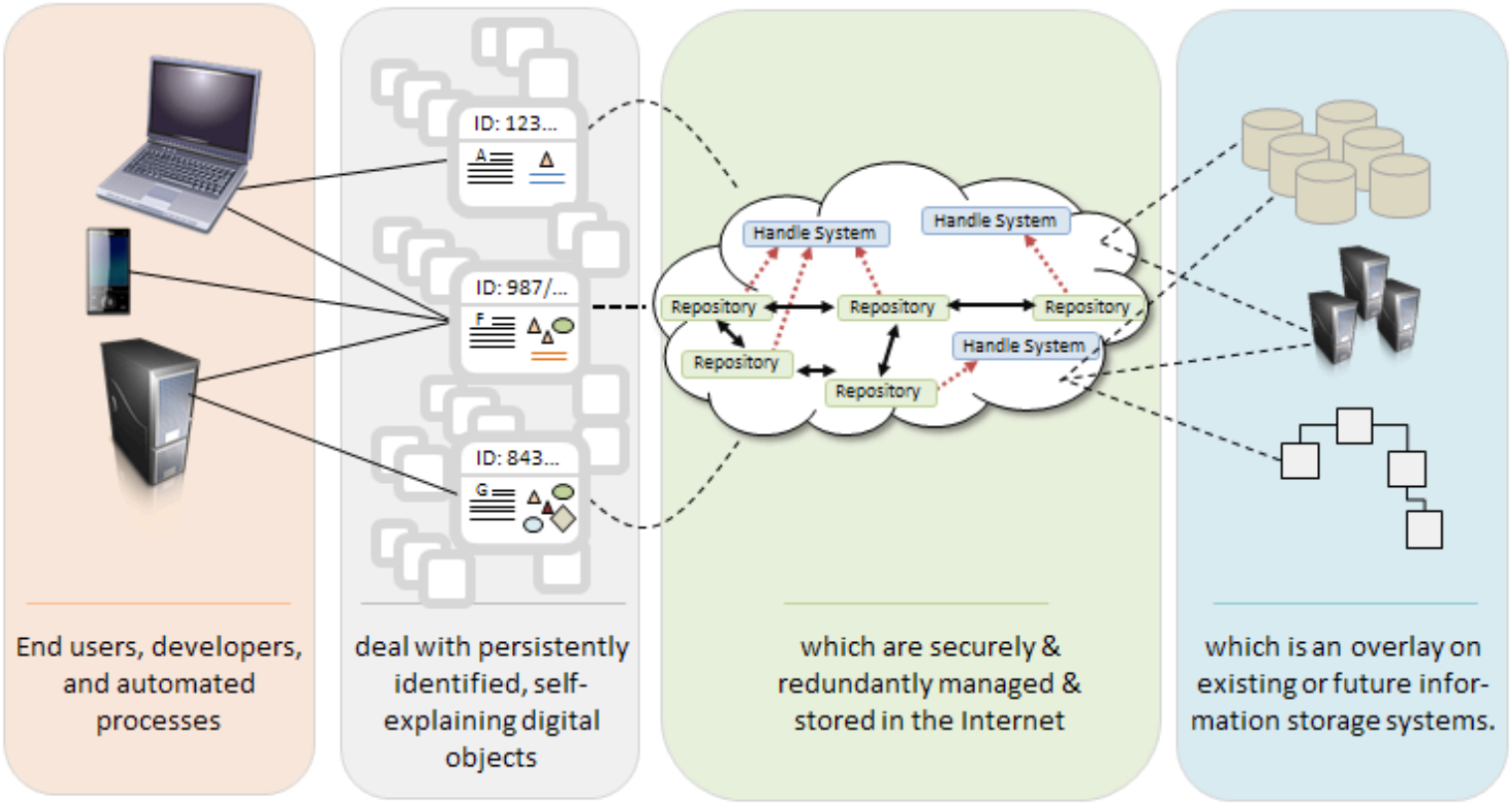
Irino, T; Tada, R (2009): Chemical and mineral compositions of sediments from ODP Site 127-797. Geological Institute, University of Tokyo.http://dx.doi.org/10.1594/PANGAEA.726855

Geofon operator (2009): GEFON event gfz2009kciu (NW Balkan Region). GeoForschungsZentrum Potsdam (GFZ).http://dx.doi.org/10.1594/GFZ.GEOFON.gfz2009kciu

Denhard, Michael (2009): dphase_mpeps: MicroPEPS LAF-Ensemble run by DWD for the MAP D-PHASE project. World Data Center for Climate http://dx.doi.org/10.1594/WDCC/dphase_mpeps

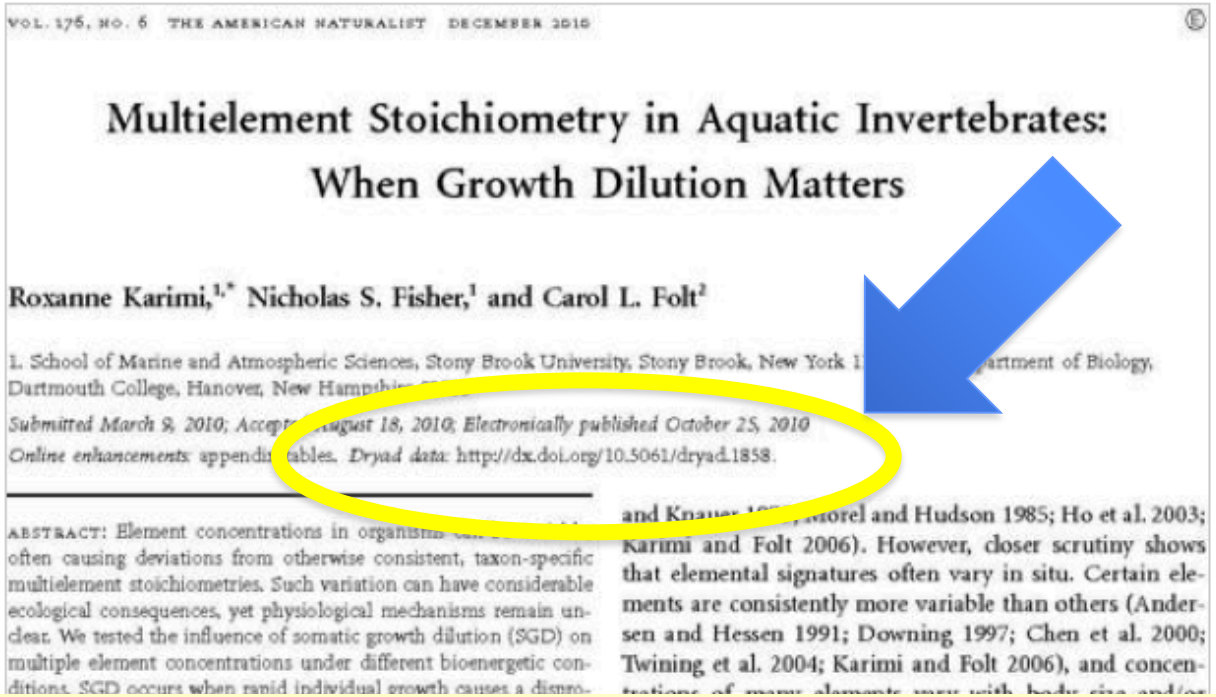
7 – D.O.I. para Dados

Arquitetura



DOI Handbook: Acesso disponível em: https://www.doi.org/doi_handbook/5_Applications.html

7 – D.O.I. para Dados – Citação



VOL. 176, NO. 6 THE AMERICAN NATURALIST DECEMBER 2010

Multielement Stoichiometry in Aquatic Invertebrates: When Growth Dilution Matters

Roxanne Karimi,[1,*] Nicholas S. Fisher,[1] and Carol L. Folt[2]

1. School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York 1
artment of Biology, Dartmouth College, Hanover, New Hampshi

Submitted March 9, 2010; Accept ugust 18, 2010; Electronically published October 25, 2010

Online enhancements: appendi ables. *Dryad data:* http://dx.doi.org/10.5061/dryad.1858.

ABSTRACT: Element concentrations in organisms
often causing deviations from otherwise consistent, taxon-specific multielement stoichiometries. Such variation can have considerable ecological consequences, yet physiological mechanisms remain unclear. We tested the influence of somatic growth dilution (SGD) on multiple element concentrations under different bioenergetic conditions. SGD occurs when rapid individual growth causes a dispro-

and Knauer 1
Morel and Hudson 1985; Ho et al. 2003; Karimi and Folt 2006). However, closer scrutiny shows that elemental signatures often vary in situ. Certain elements are consistently more variable than others (Andersen and Hessen 1991; Downing 1997; Chen et al. 2000; Twining et al. 2004; Karimi and Folt 2006), and concen-
trations of many elements may with body size and/or

When using this data, please cite the original article:

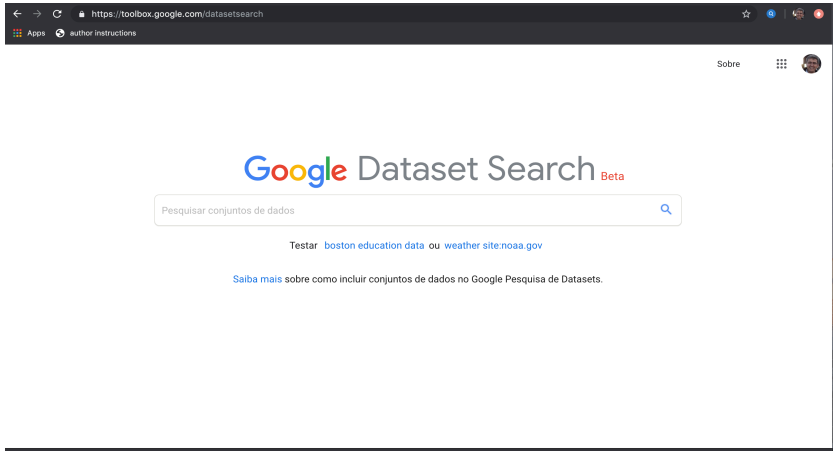
> Ally D, Ritland K, Otto SP (2008) Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in Populus tremuloides. Molecular Ecology 17(22): 4897-4911. doi:10.1111/j.1365-294X.2008.03962.x

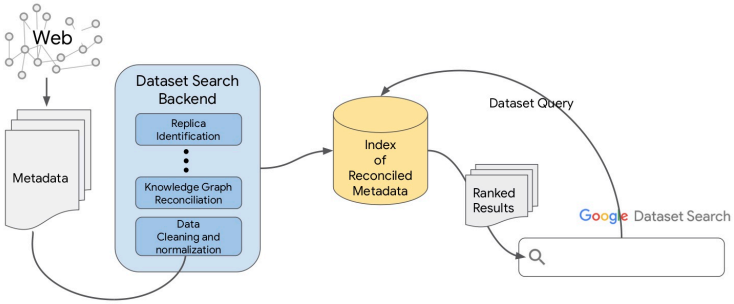
Additionally, please cite the Dryad data package:

> Ally D, Ritland K, Otto SP (2008) Data from: Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in Populus tremuloides. Dryad Digital Repository. doi:10.5061/dryad.7898

7 – Exemplo Publicação de Dados

Google Data Search
https://toolbox.google.com/datasetsearch

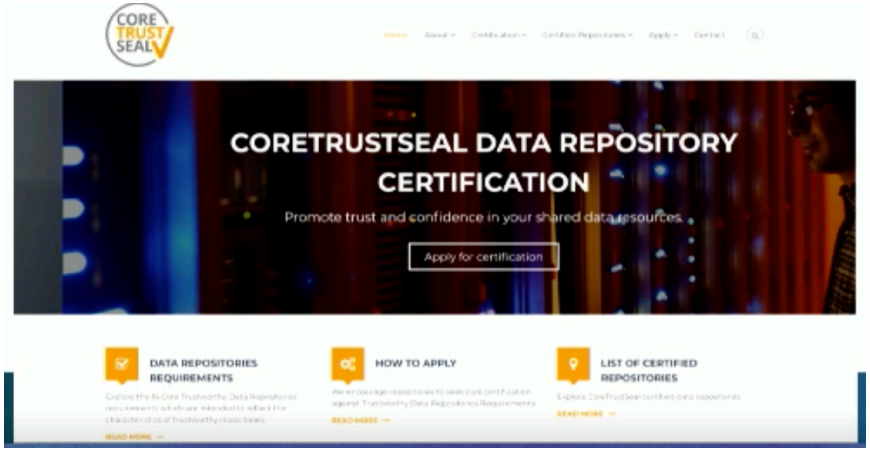


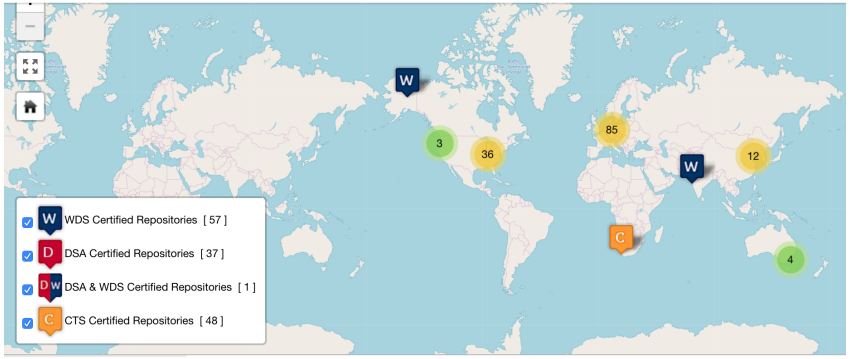


Natasha Noy and Matthew Burgess and Dan Brickley , Google Dataset Search: Building a search engine for datasets in an open Web ecosystem, WebConf 2019. available at: https://ai.google/research/pubs/pub47845

Confiabilidade para reuso dos dados

- Se alguém entregar seus dados científicos, o que é necessário para convencê-lo que os dados estão corretos ?
- Se requer um sistema complexo para executá-lo e que não tem acesso, o que precisa para confiar dos dados ? Você conhece quais são as suposições e dependências existentes ?
- Quanto você poderá confiar que os mesmo dado estará disponível por um longo período de tempo ?





Boas Práticas na Gestão de Dados

Integrar → **Analisar** → Planejar

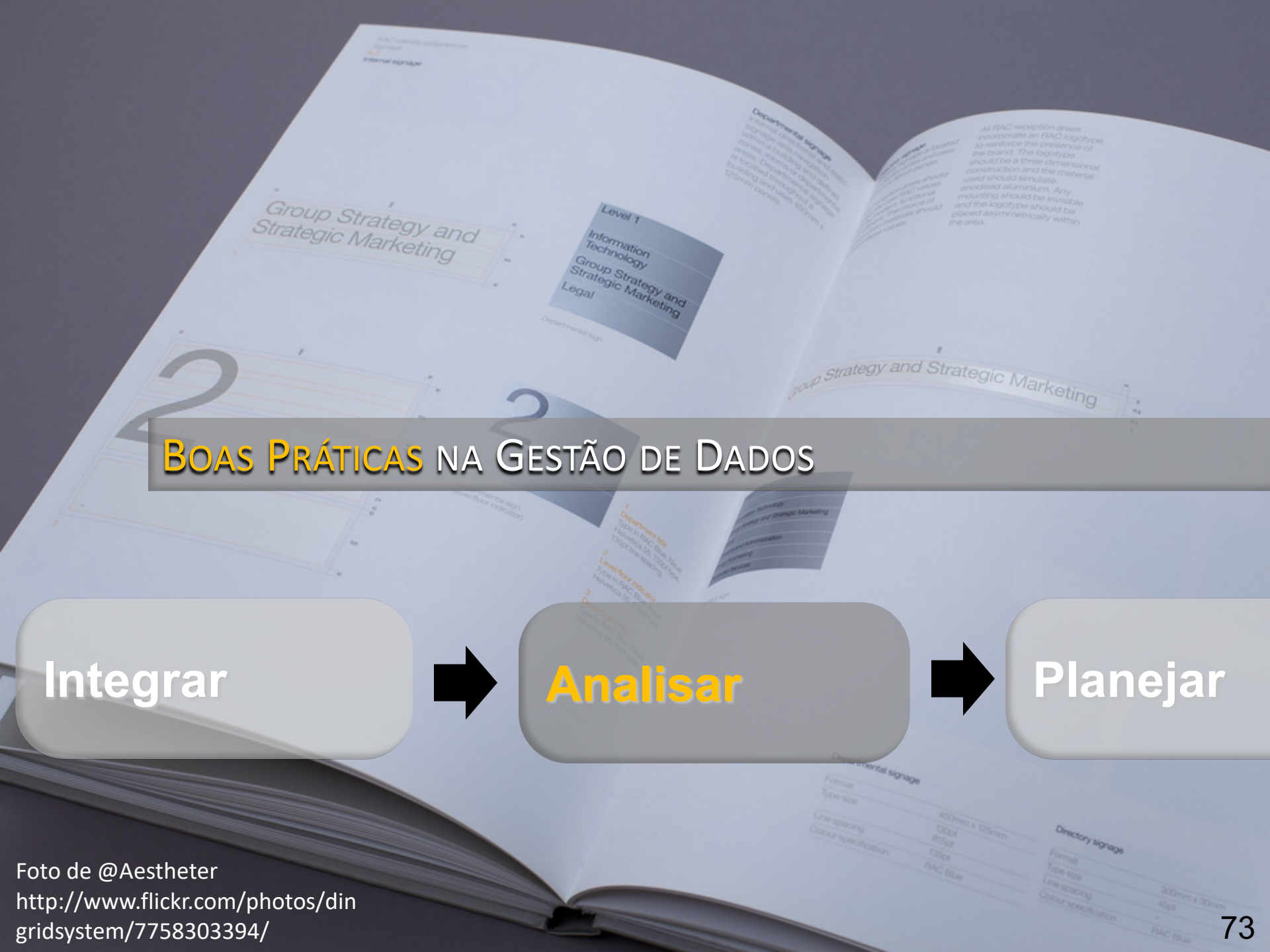
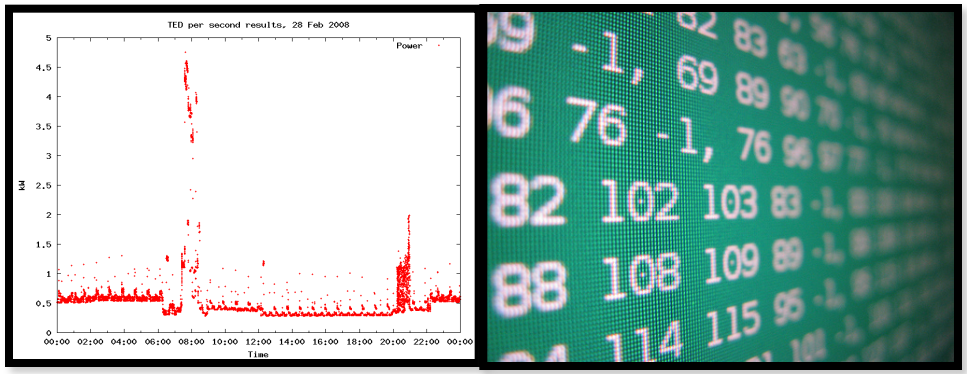


Foto de @Aestheter
http://www.flickr.com/photos/dingridsystem/7758303394/

8 - Análise

a) Utilizar os dados para análises que atendam os objetivos do projeto.

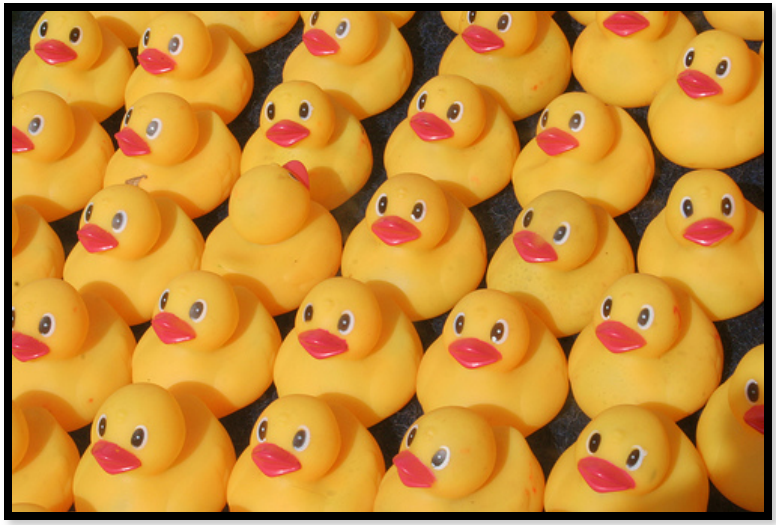
- Muitas ferramentas de software já estão disponíveis para apoiar as atividades de exploração, análise e visualização dos dados.
- Data Analyses
  - Conduzida por computadores, grid ou *cloud computing*;
  - Estatísticas, execução de modelos, estimativa de parâmetros, gráficos e plotagens.



8 - Análise

b) Reprodutibilidade

- **A reprodutibilidade é a chave para os métodos científicos;**
- **Processos complexos são mais difíceis de serem reproduzidos;**
- **Boa documentação é essencial para a reprodutibilidade;**



8 - Análise

Fluxos de trabalho (*workflow*)

- A formalização dos metadados do processo;
- Descrição precisa do procedimento científico;
- Três componentes:
  - Inputs: Informação e/ou material necessário;
  - Outputs: Informação ou material produzido e potencialmente utilizado como input em outros passos;
  - Regras de transformação/algoritmos.

8 - Análise

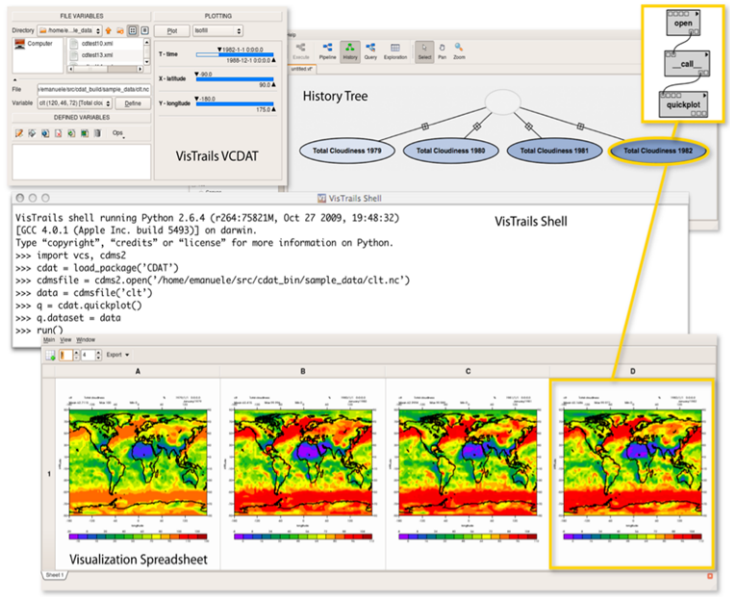
Boas práticas

- Os cientistas devem documentar os fluxos de trabalho usados na criação de resultados:
  - Proveniência dos dados;
  - Análises e parâmetros utilizados;
  - Conexões entre análises por meio dos inputs (entradas e outputs (saídas).
- A documentação pode ser informal (ex: flowcharts, *commented scripts*) ou formal (ex: Kepler, Vistrails).

8 - Análise

- Workflow Formal

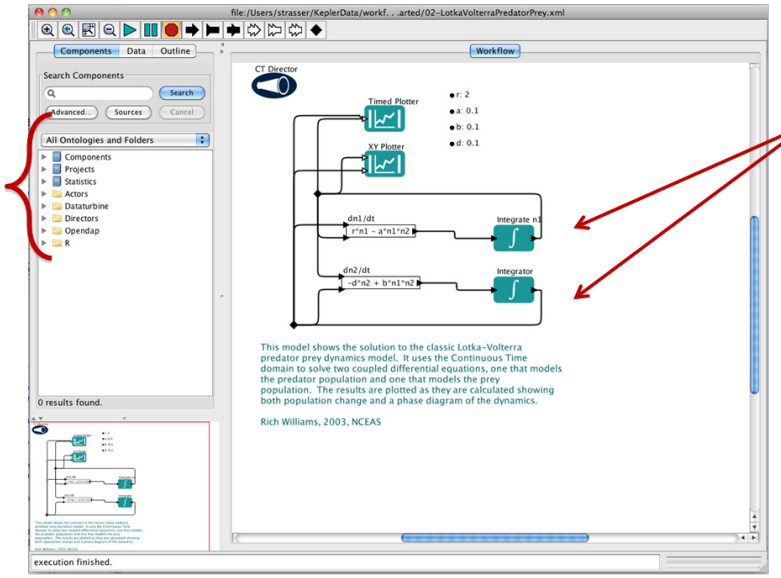
VisTrails



www.vistrails.org

Drag & drop components from this list

Kepler Software



Actors in workflow

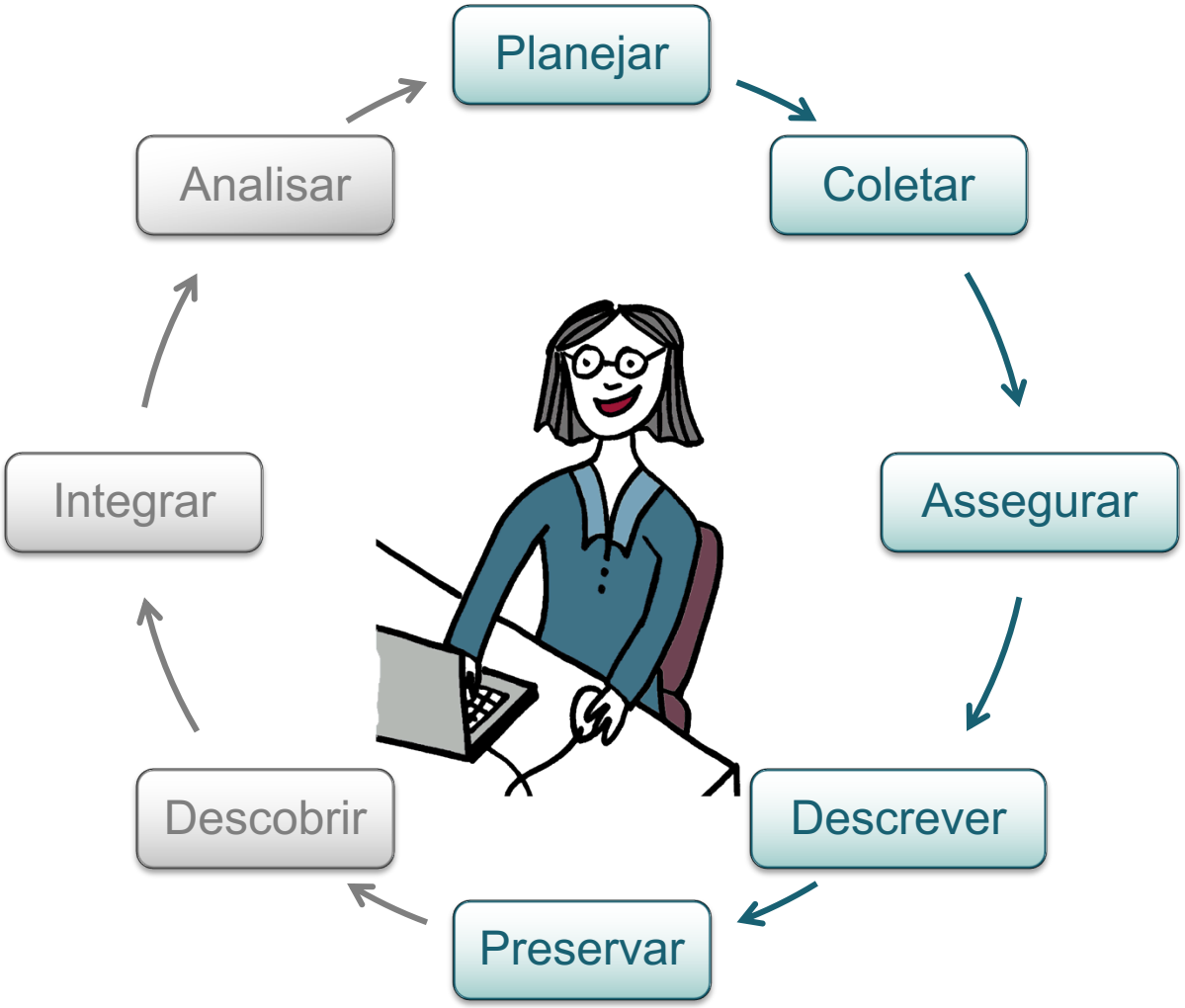
kepler-project.org

Software - Próxima fronteira

- Similar ao processo de publicação dos Dados
- O software que Analisa os dados precisa ser preservado para efeitos e reprodutibilidade
- Considerações: https://www.usgs.gov/media/files/usgs-software-planning-checklist
  - Categorias de disponibilização – Informal e formal
  - Formal: Disclaimers (Provisório/Aprovado)
  - 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

USGS science for a changing world

Gestão de Dados



Exercício - Desafio

1- Considere um domínio de aplicação envolvendo dados de seu interesse. Defina um plano de gestão de dados considerando a ferramenta dmp.cdlib.org.

2- Para esse domínio selecione um arquivo em formato csv, com pelo menos 2 atributos. Defina os metadados para esse arquivo utilizando as propriedades recomendadas do Google Search.

Referências:

- Diretrizes para DataSets https://developers.google.com/search/docs/data-types/dataset

- Natasha Noy and Matthew Burgess and Dan Brickley , Google Dataset Search: Building a search engine for datasets in an open Web ecosystem, WebConf 2019. available at: https://ai.google/research/pubs/pub47845

Conceitos de Big Data

PCS5787 – Ciência dos Dados
Prof. Dr. Pedro Luiz Pizzigatti Corrêa
24 de Setembro de 2020