

A vida secreta dos dados

Manejo, Visualização e Compartilhamento de Dados

Nicholas A. C. Marino

nac.marino@gmail.com

github.com/nacmarino/compartilhaR



Existem algumas coisas sobre as quais falamos muito na ciência...

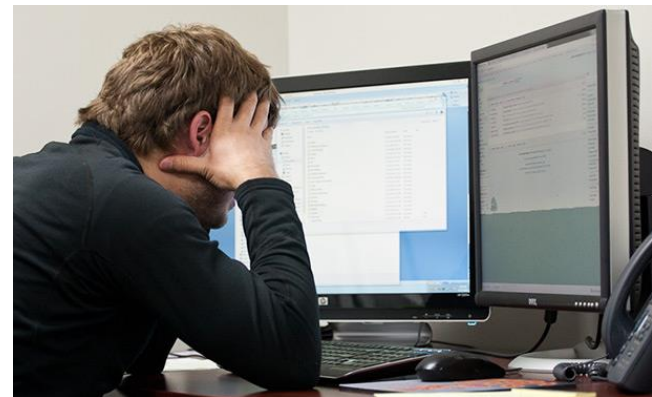
Idéias



Trabalhos e coletas de campo

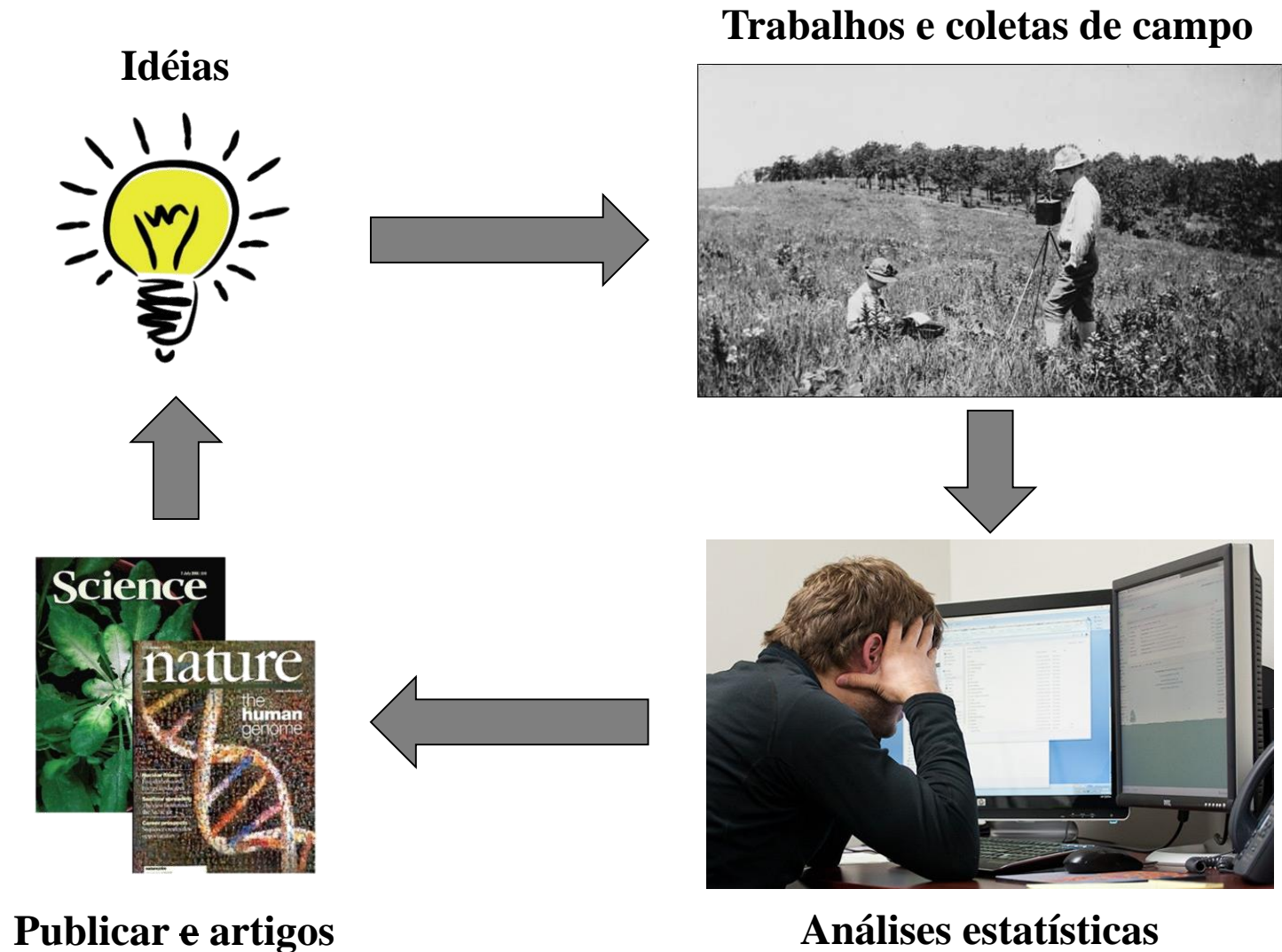


Publicar e artigos



Análises estatísticas

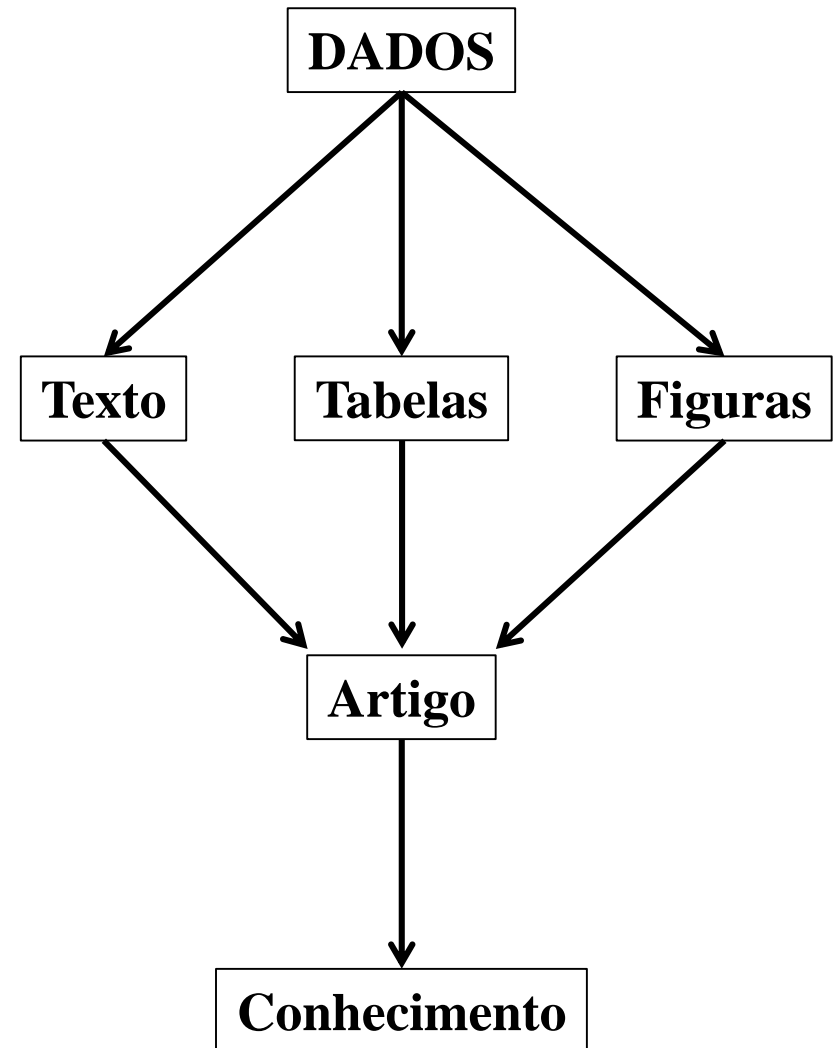
...todas essas coisas estão bastante relacionadas...



...mas falhamos em dar importância aquilo que realmente une tudo isso: os dados.



Também é comum achar que entre ter o dado em mãos e gerar um conhecimento a partir deles é tarefa fácil e rápida, mas...



Existe um precipício enorme entre os dados e o que podemos gerar a partir deles



Existe um precipício enorme entre os dados e o que podemos gerar a partir deles

1. Falha em registrar informações de forma a facilitar seu uso;



Existe um precipício enorme entre os dados e o que podemos gerar a partir deles

1. Falha em registrar informações de forma a facilitar seu uso;
2. Dificuldade de encontrar uma informação da forma como você havia imaginado;



Existe um precipício enorme entre os dados e o que podemos gerar a partir deles

1. Falha em registrar informações de forma a facilitar seu uso;
2. Dificuldade de encontrar uma informação da forma como você havia imaginado;
3. Falta de consistência na forma como uma informação é registrada e apresentada;



Existe um precipício enorme entre os dados e o que podemos gerar a partir deles



1. Falha em registrar informações de forma a facilitar seu uso;
2. Dificuldade de encontrar uma informação da forma como você havia imaginado;
3. Falta de consistência na forma como uma informação é registrada e apresentada;
4. Desorganização no processamento e armazenamento dos dados;

Existe um precipício enorme entre os dados e o que podemos gerar a partir deles



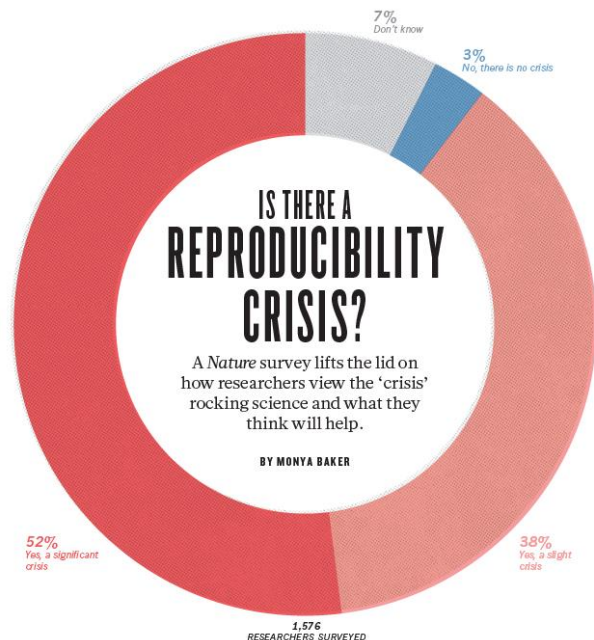
1. Falha em registrar informações de forma a facilitar seu uso;
2. Dificuldade de encontrar uma informação da forma como você havia imaginado;
3. Falta de consistência na forma como uma informação é registrada e apresentada;
4. Desorganização no processamento e armazenamento dos dados;
5. Má interpretação do significado dos dados.

Existe um precipício enorme entre os dados e o que podemos gerar a partir deles – mas tem jeito!



1. Falha em registrar informações de forma a facilitar seu uso;
2. Dificuldade de encontrar uma informação da forma como você havia imaginado;
3. Falta de consistência na forma como uma informação é registrada e apresentada;
4. Desorganização no processamento e armazenamento dos dados;
5. Má interpretação do significado dos dados.

O problema é que não falamos desses jeitos, tampouco do que nos abster pode causar...



NATURE | NEWS

Scientists losing data at a rapid rate

Decline can mean 80% of data are unavailable after 20 years.

Elizabeth Gibney & Richard Van Noorden

PLOS | COMPUTATIONAL BIOLOGY
EDUCATION

Unmet needs for analyzing biological big data: A survey of 704 NSF principal investigators

Lindsay Barone^{*}, Jason Williams^{*}, David Micklos^{*}

PeerJ

Lack of quantitative training among early-career ecologists: a survey of the problem and potential solutions

Frédéric Barraquand^{1,11}, Thomas H.G. Ezard^{2,11}, Peter S. Jørgensen^{3,11}, Naupaka Zimmerman^{4,11}, Scott Chamberlain⁵, Roberto Salguero-Gómez^{6,7,11}, Timothy J. Curran^{8,11} and Timothée Poisot^{9,10,11}

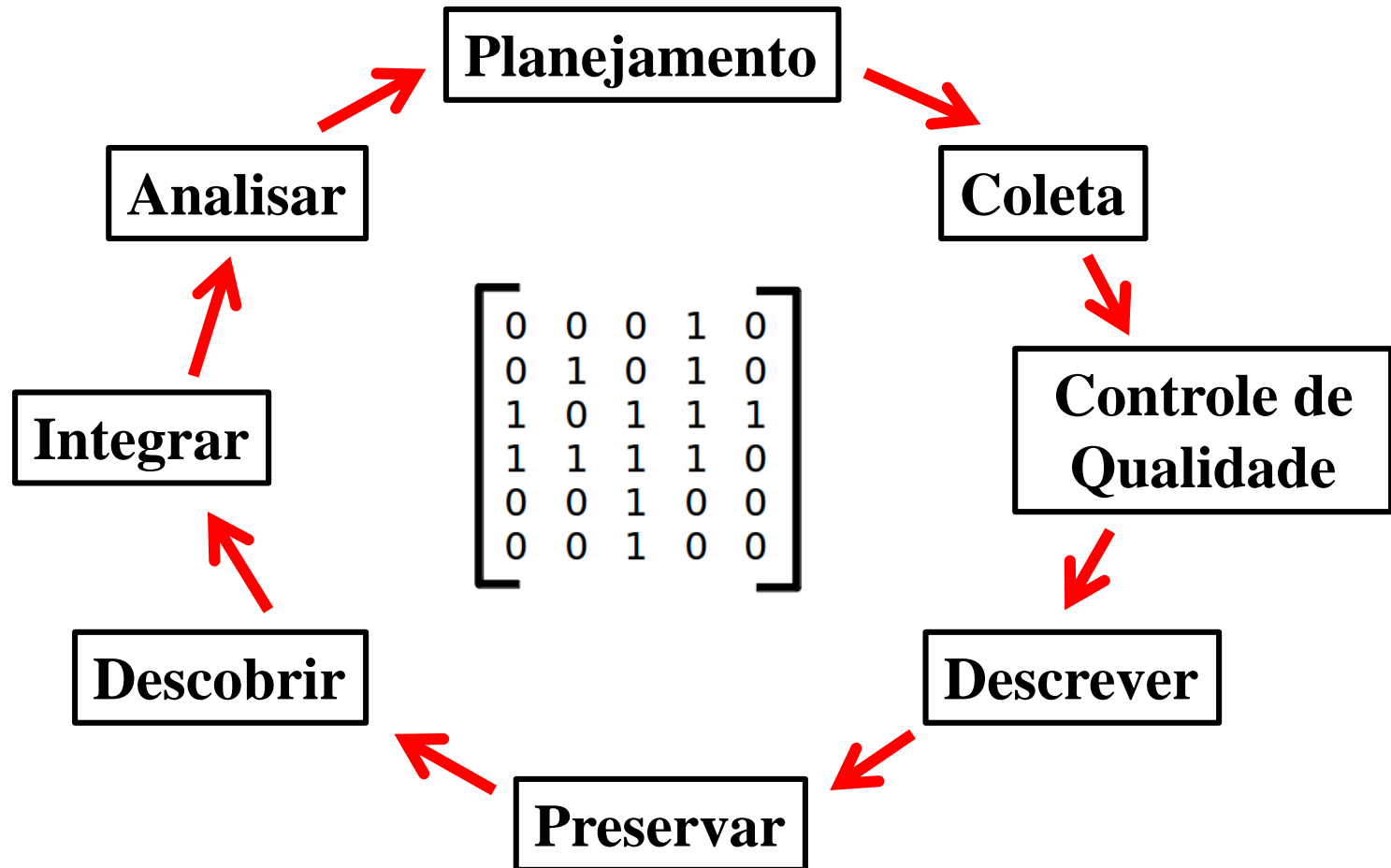
ORIGINAL RESEARCH

WILEY Ecology and Evolution

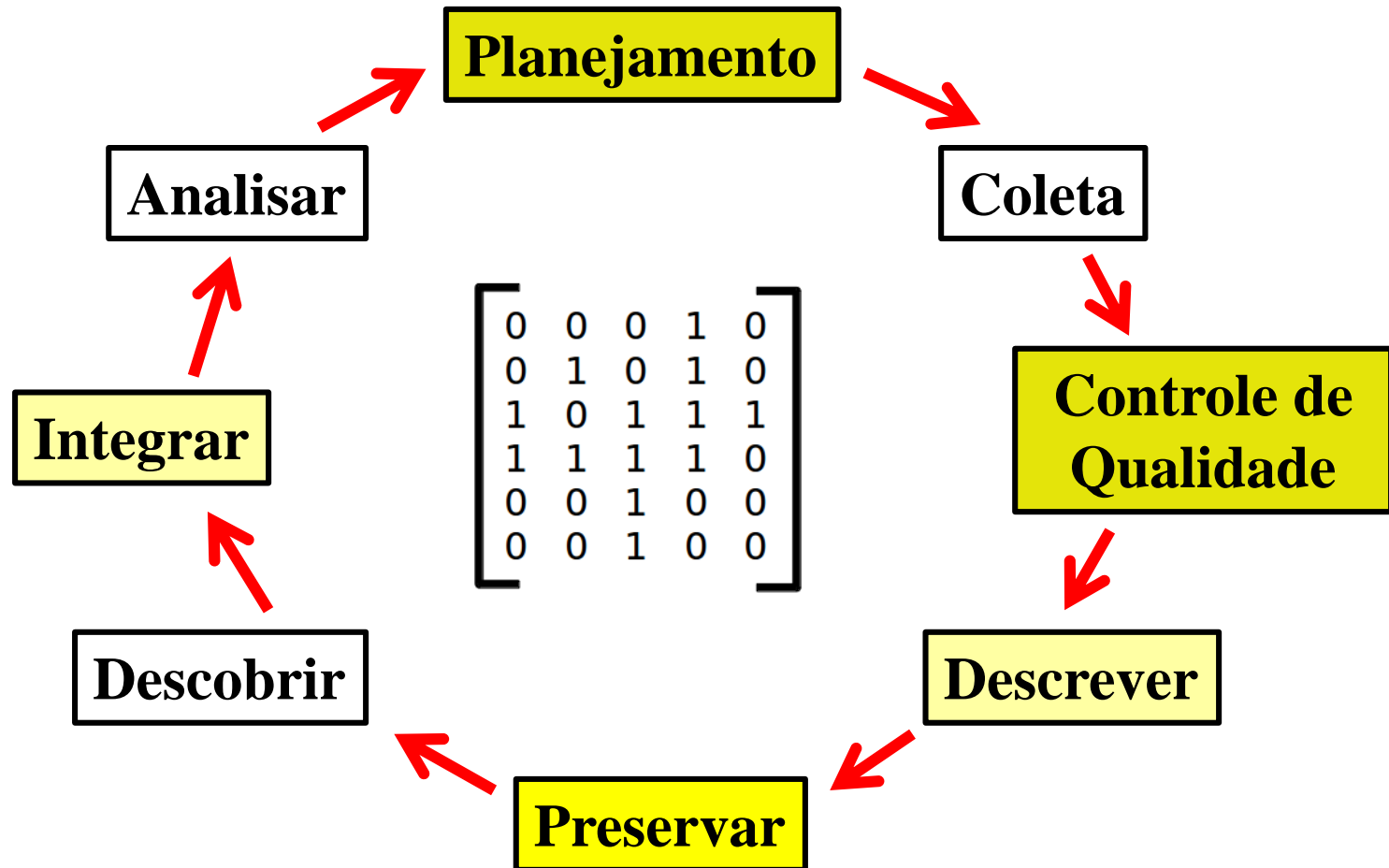
Temporal degradation of data limits biodiversity research

Geiziane Tessarolo^{1,2} | Richard Ladle^{3,4} | Thiago Rangel¹ | Joaquin Hortal^{1,5}

Uma solução é reconhecer que todo o dado possui um ciclo de vida, e atuarmos ativamente em cada uma de suas etapas



Existem cinco aspectos particularmente marcantes deste ciclo que normalmente não damos muita atenção



Por que esses cinco aspectos?

Como reunir o que não foi
pensado para tal?



Analisar

Integrar

Descobrir

Planejamento



Leva tempo, que nós não
queremos perder

Coleta

Controle de
Qualidade



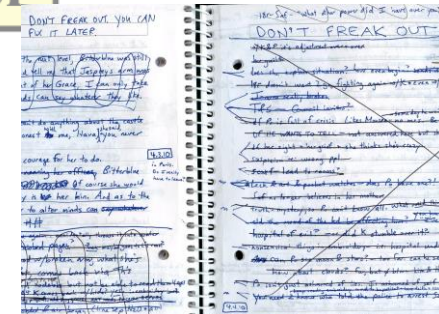
Descrever



Preservar

“Em algum desses pen
drives tem uma cópia!”

Relatamos mal o que fazemos, logo
não mantemos registros



Exercício #1

- **Como o investimento em pesquisa se traduz em produção científica?**
- Extrair e registrar as informações que você julgar necessária para responder essa pergunta dos relatórios finais dos projetos financiados pelo CNPq.
 - ✓ Tarefa em trios – 20 minutos;
 - ✓ Ache um novo trio e compare de que forma vocês fizeram isso – 10 minutos;
 - ✓ Retorne ao seu trio original e troque experiências – 10 minutos.

**Algumas dicas para o planejamento e
a etapa inicial do controle de
qualidade dos dados**

Cada linha corresponde à uma unidade observacional, e cada coluna à uma variável que a descreva

	PONTO 1		PONTO 2		PONTO 3	
variavel	chuva	estiagem	chuva	estiagem	chuva	estiagem
v1	presente	presente	ausente	ausente	presente	ausente
v2	7.2	4.5	3.4	3.2	6.8	3.6
v3	F	V	V	V	F	V
v4	-72.4;-23.4	-72.4;-23.4	-73.2;-22.1	-73.2;-22.1	-71.5;-20.5	-71.5;-20.5



site	estacao	v1	v2	v3	v4a	v4b
Ponto 1	chuva	presente	7.2	F	-72.4	-23.4
Ponto 1	estiagem	presente	4.5	V	-72.4	-23.4
Ponto 2	chuva	ausente	3.4	V	-73.2	-22.1
Ponto 2	estiagem	ausente	3.2	V	-73.2	-22.1
Ponto 3	chuva	presente	6.8	F	-71.5	-20.5
Ponto 3	estiagem	ausente	3.6	V	-71.5	-20.5



Cada observação deve receber um código de identificação único, que não se repete em nenhuma outra linha

ID	site	estacao	v1	v2	v3	v4.long	v4.lat
P1C	Ponto 1	chuva	presente	7.2	F	-72.4	-23.4
P1E	Ponto 1	estiagem	presente	4.5	V	-72.4	-23.4
P2C	Ponto 2	chuva	ausente	3.4	V	-73.2	-22.1
P2E	Ponto 2	estiagem	ausente	3.2	V	-73.2	-22.1
P3C	Ponto 3	chuva	presente	6.8	F	-71.5	-20.5
P3E	Ponto 3	estiagem	ausente	3.6	V	-71.5	-20.5

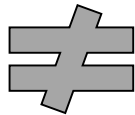
Na medida do possível, registre o que irá em cada coluna de uma tabela e que valores ela deve receber

- A ideia aqui é fazer o exercício de projetar a estrutura da tabela e te ajudar, no futuro, qual foi a sua linha de raciocínio ao definir cada variável.

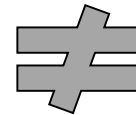
coluna	descrição	tipo	valores
Site	Nome do local onde a amostragem foi feita	Categórico	Rio de Janeiro, São Paulo ou Minas Gerais
Altitude	Altitude do local onde a amostragem foi feita, em metros	Contínuo	Números de 0 a 2000
Distúrbio	Valor que descreve se a localidade de coleta está sob efeito de algum distúrbio	Lógico	VERDADEIRO ou FALSO

Os valores atribuídos a cada coluna devem ser consistentes tanto no formato quanto na escrita

Presente

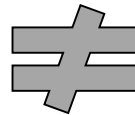


presente



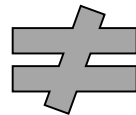
1

Verão + Chuva

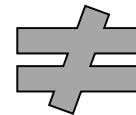


Chuva + Verão

Homo sapiens



Homo.sapiens



Homo sapiens

Valores não registrados ou faltantes devem receber o código 'NA'

ID	V1	V2	V3	V4	V5	V6
AN1	9	nd	5	45	21	10
AN2	-	10	30	3	na	25
CD1	20	40	Falta		15	0



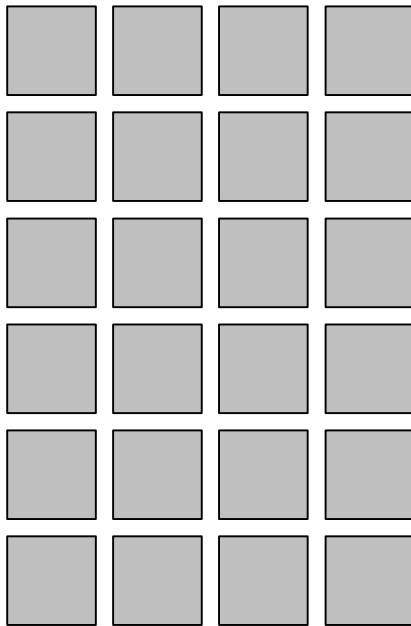
←
**Dado perdido, faltante,
não registrado,...**

ID	V1	V2	V3	V4	V5	V6
AN1	9	NA	5	45	21	10
AN2	NA	10	30	3	NA	25
CD1	20	40	NA	NA	15	NA

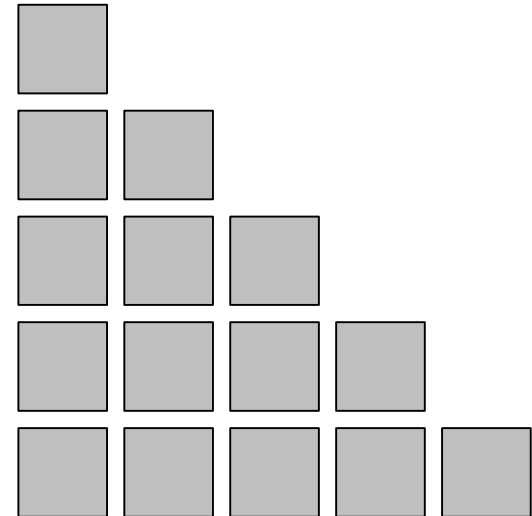


Dados com natureza e estrutura diferentes devem ficar em tabelas diferentes

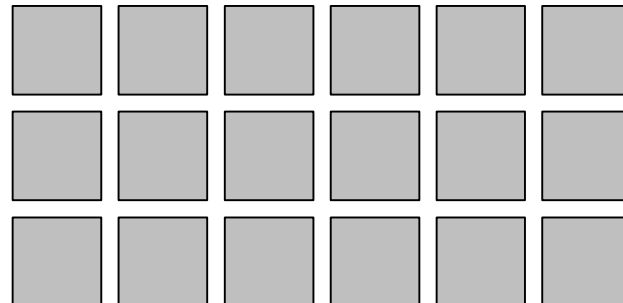
Características de cada indivíduo



Distância filogenética entre as espécies

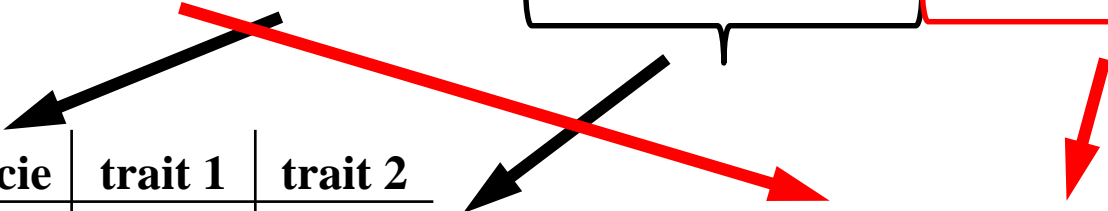


Composição da comunidade



Na medida do possível, não se repita!

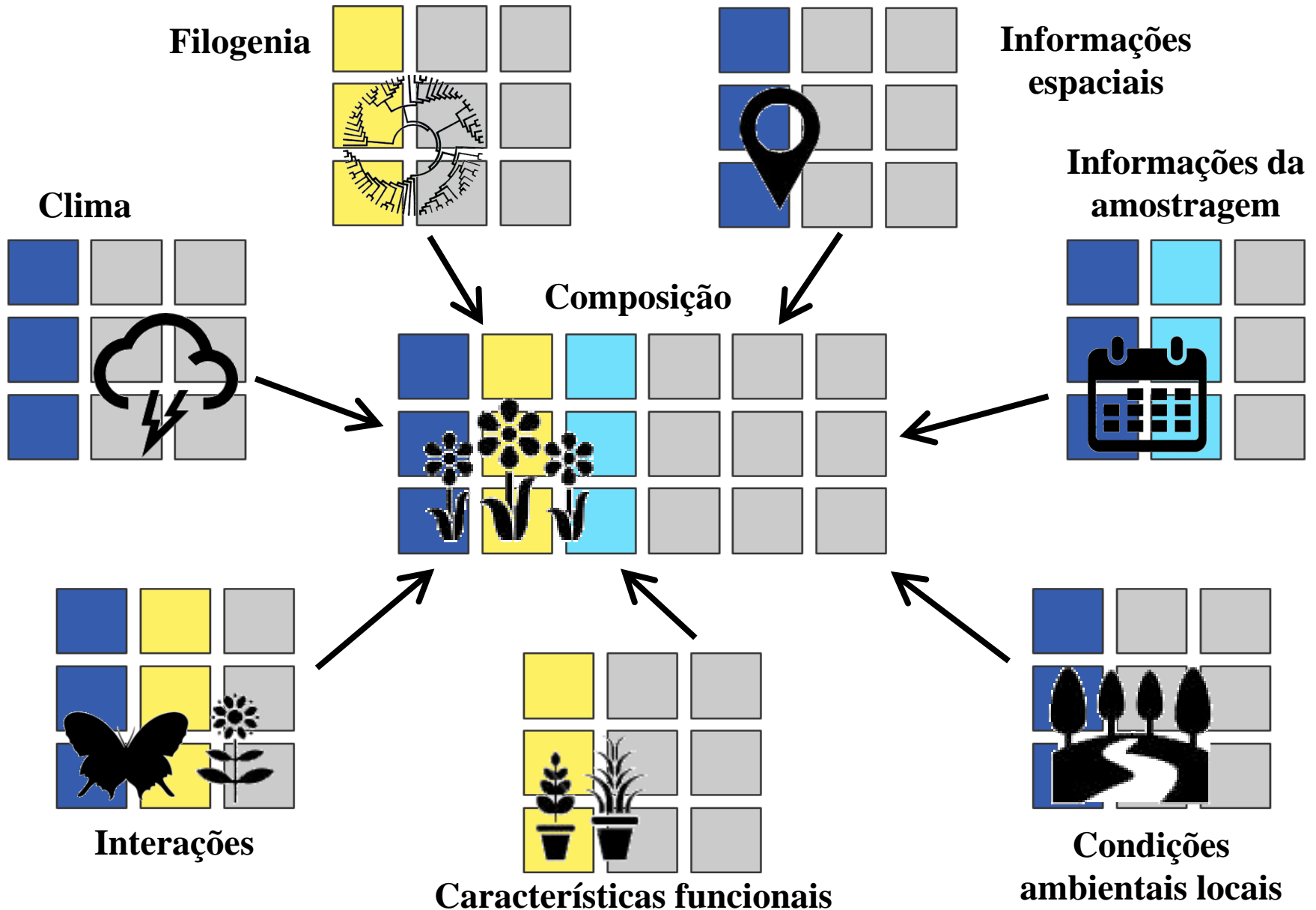
ID	site	especie	abund	trait 1	trait 2	altitude	vazão
1	site1	sp2	20	A	2.1	560	3.4
2	site1	sp4	5	B	4.3	560	3.4
3	site1	sp5	30	C	1.0	560	3.4
4	site2	sp1	10	A	2.9	200	2.1
5	site2	sp2	30	A	1.8	200	2.1
6	site2	sp5	25	C	1.1	200	2.1
7	site3	sp2	15	A	2.3	940	3.5
8	site3	sp5	5	C	0.8	940	3.5



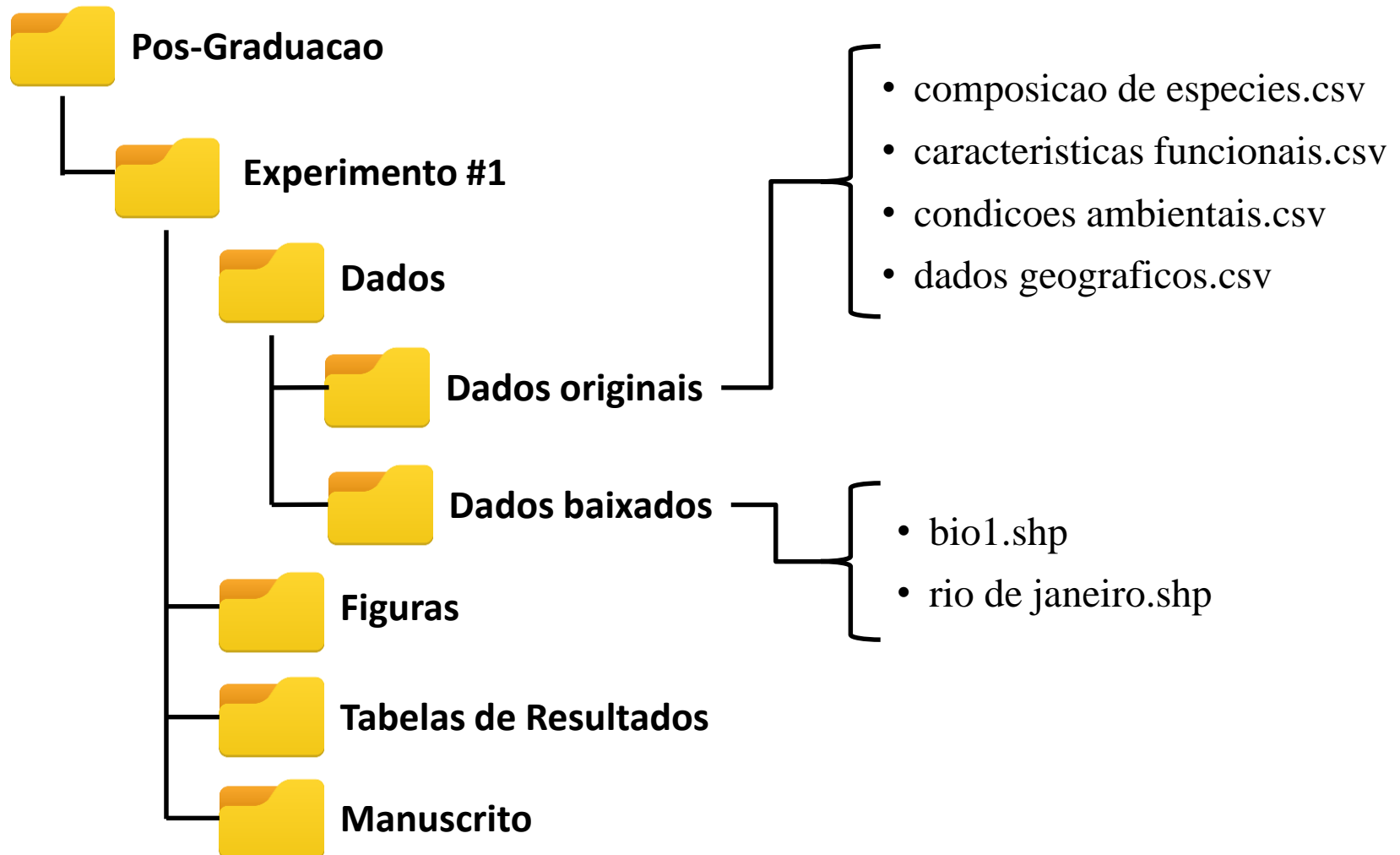
especie	trait 1	trait 2
sp2	A	2.1
sp4	B	4.3
sp5	C	1.0
sp1	A	2.9

site	altitude	vazão
site1	560	3.4
site2	200	2.1
site3	940	3.5

Utilize uma estrutura relacional entre arquivos



Organize os dados de forma intuitiva e evite múltiplas versões do mesmo arquivo



Resumindo

- Os dados são parte central de todo e qualquer trabalho, científico ou não;
- Grande parte da dificuldade que teremos (e temos) ao processar um conjunto de dados passa pela sua organização;
- Existem alguns hábitos e práticas que, se implementados, podem facilitar a rápida disponibilização e uso dos dados;
- Estas ações ajudam a sanar parte dos problemas, mas não todos! Para isso, outras ferramentas e ações são necessárias.