



Gestão e Compartilhamento de Dados

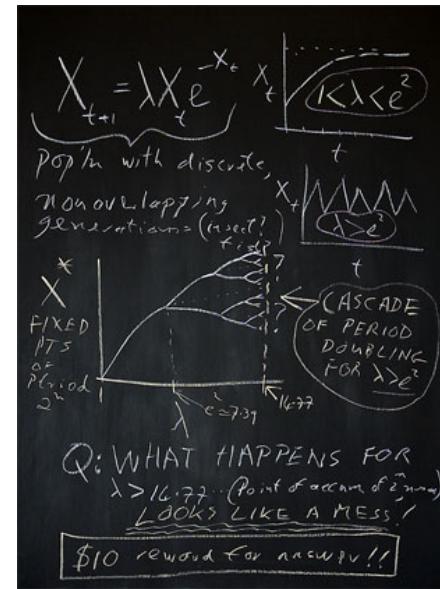
Nicholas A. C. Marino

nac.marino@gmail.com

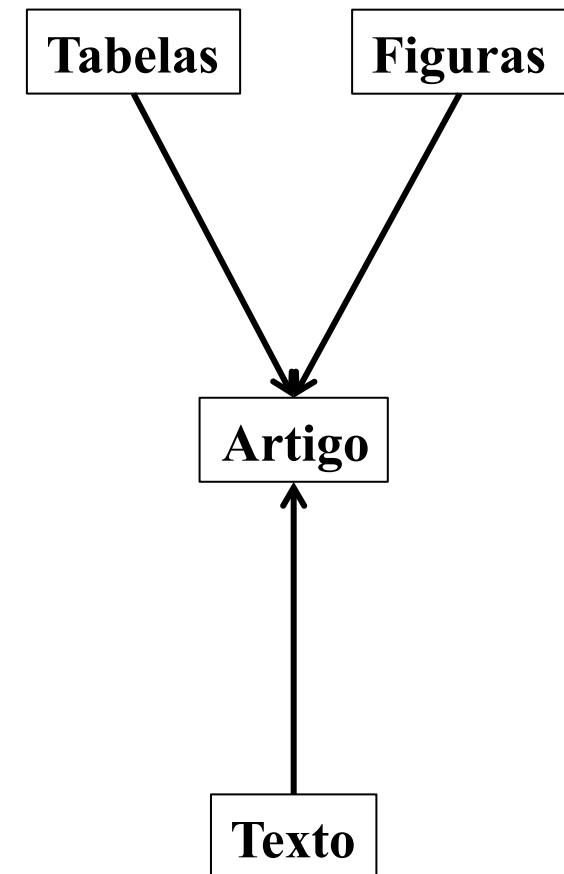
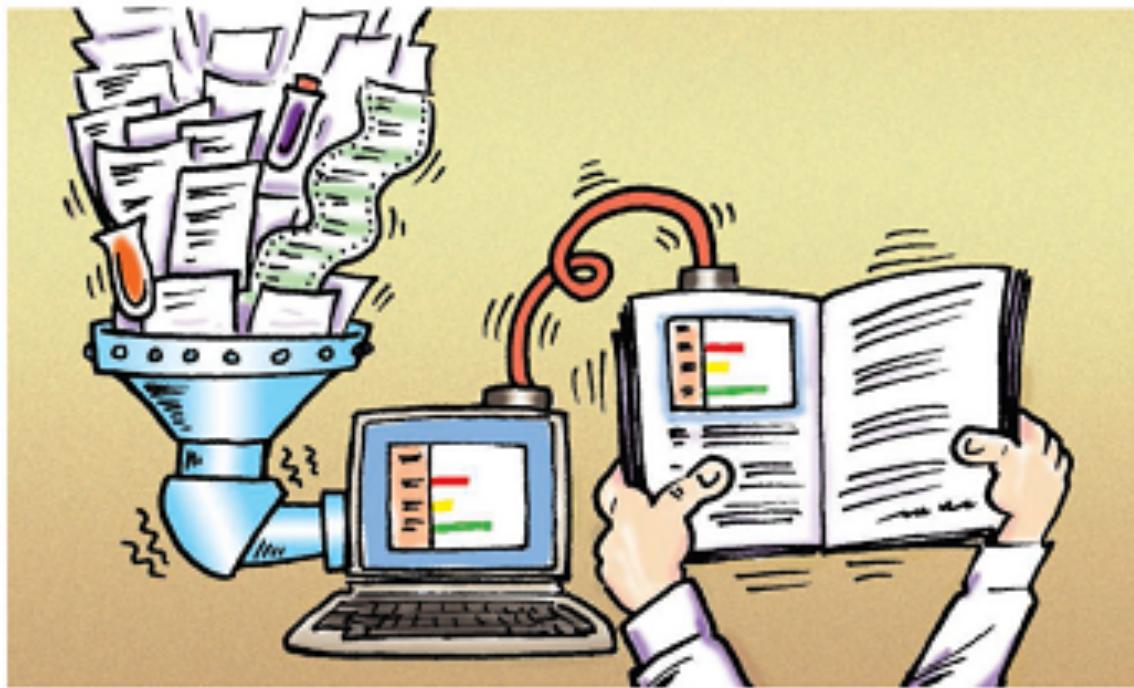
github.com/nacmarino/introducaoR

Os paradigmas de estudos científicos (na Ecologia)

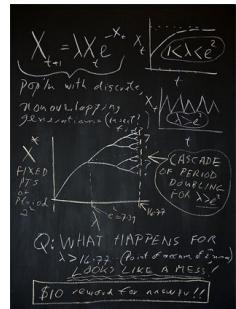
- Observacional;
- Experimental;
- Teórico;
- Computacional;



A forma clássica de escrever em ciência



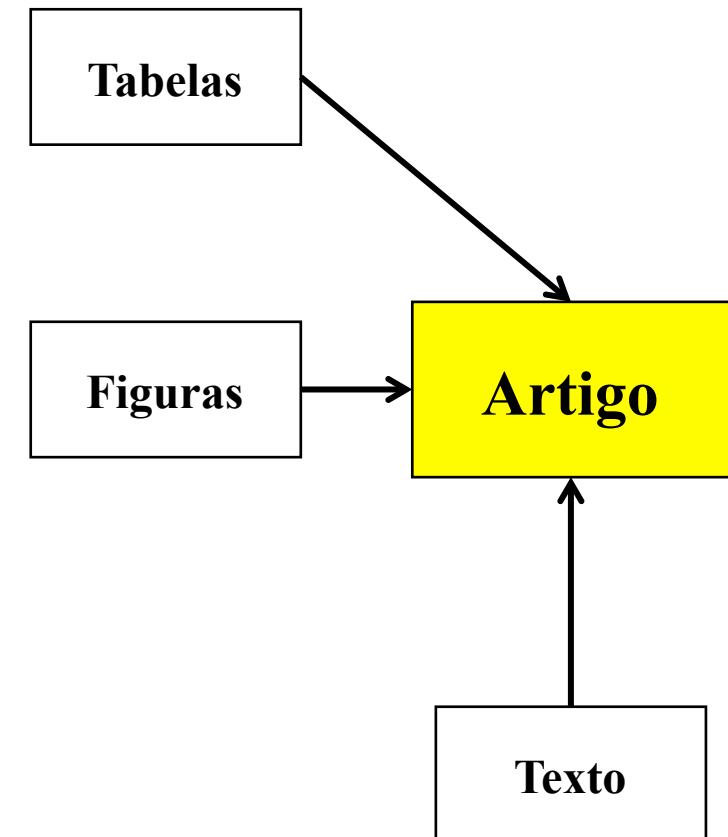
Uma nova Era para a ciência



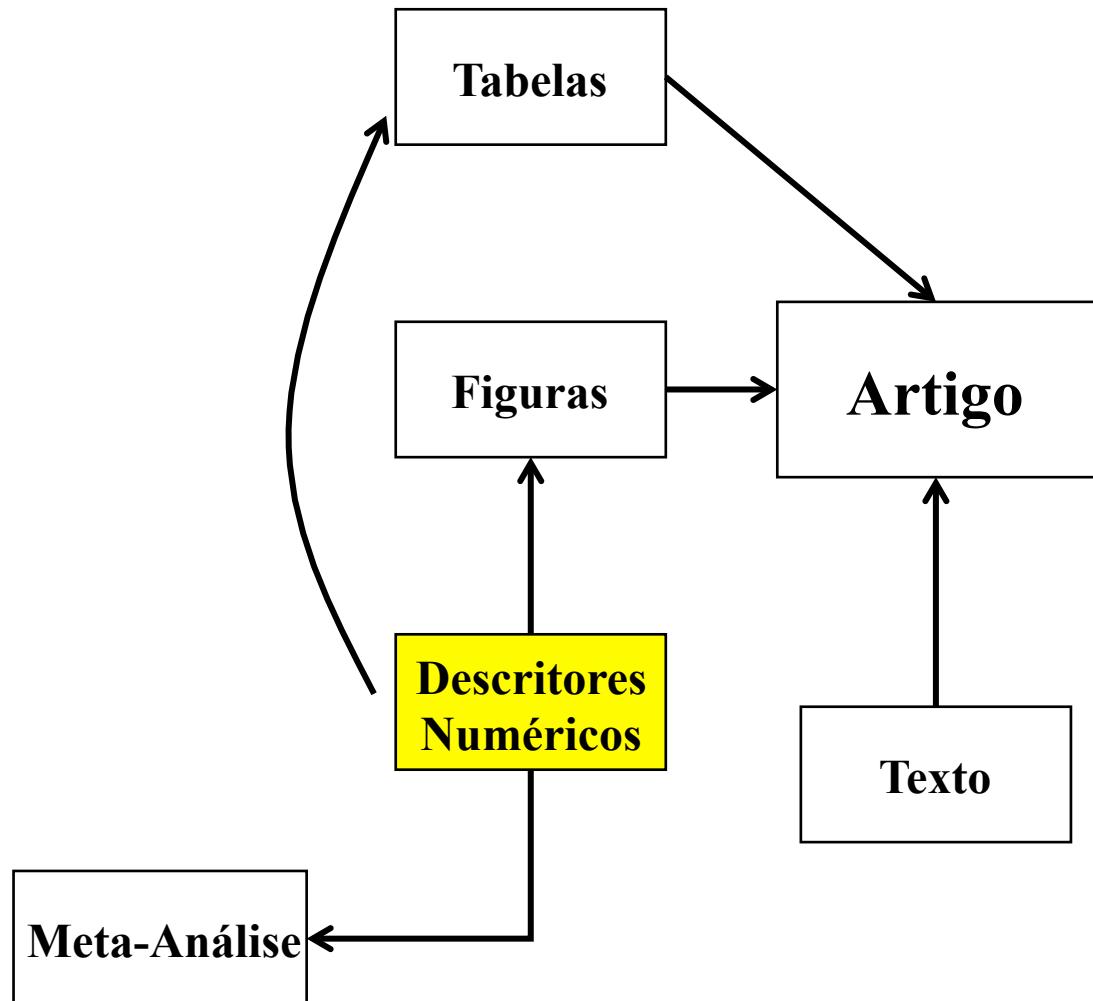
Pesquisa intensiva de dados (Data-intensive research)

- As informações estão todas disponíveis *online*;
- Os dados estão todos disponíveis *online*;
- O maior progresso na ciência ocorre pela síntese.

Uma nova Era para a ciência

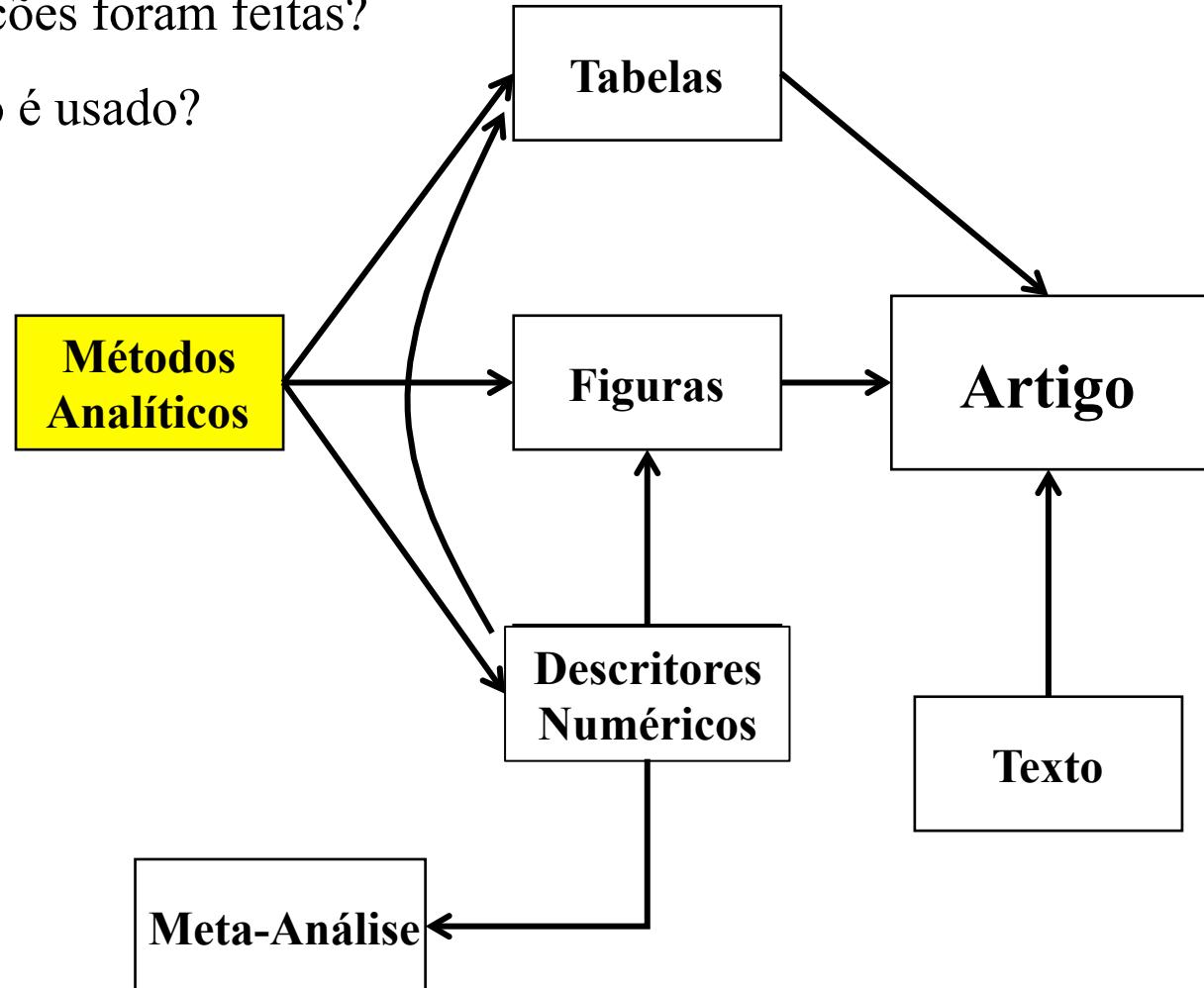


Uma nova Era para a ciência



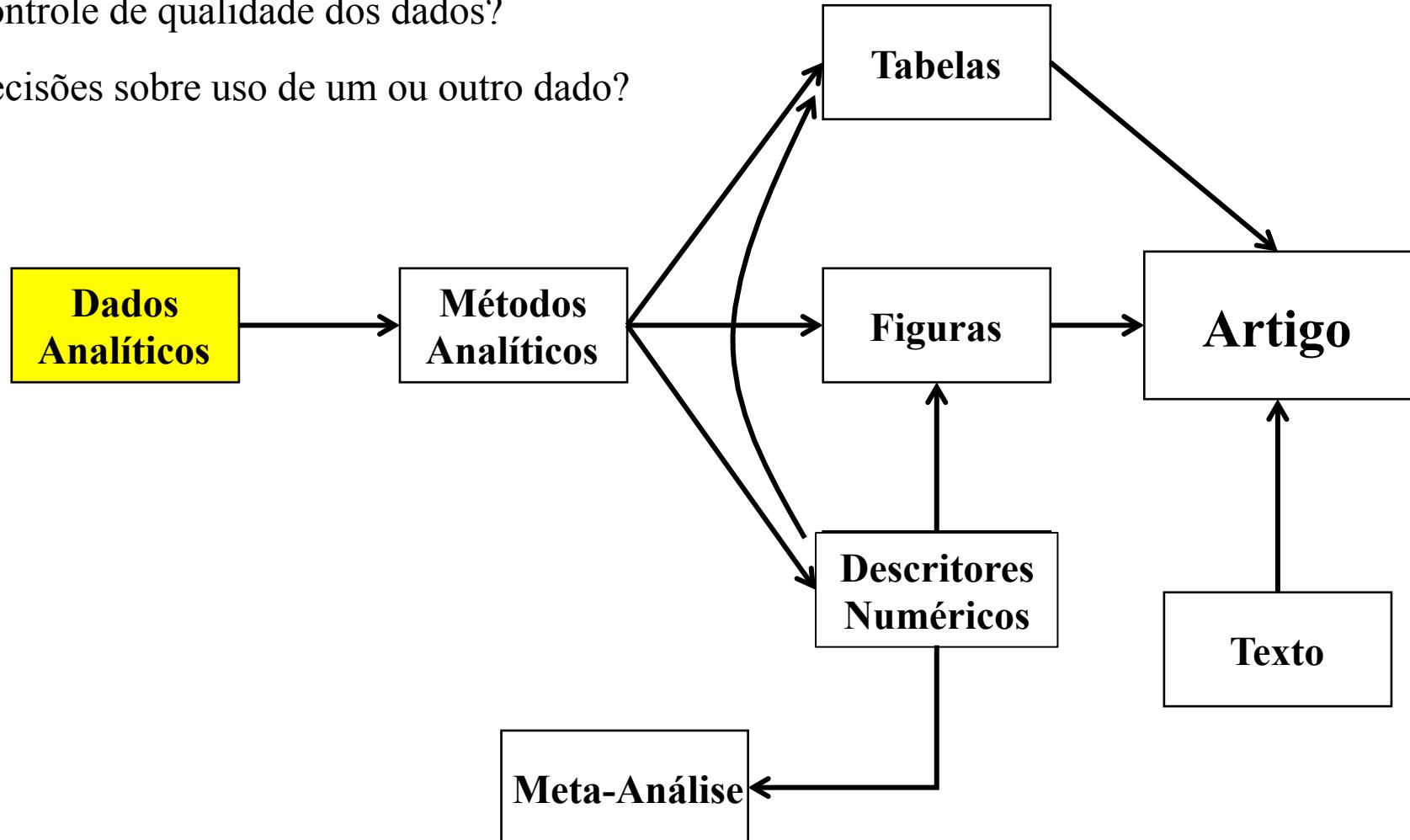
Uma nova Era para a ciência

- Como o dado foi analisado?
- Que tipo de transformações foram feitas?
- Que tipo de distribuição é usado?
- Quais os pressupostos?



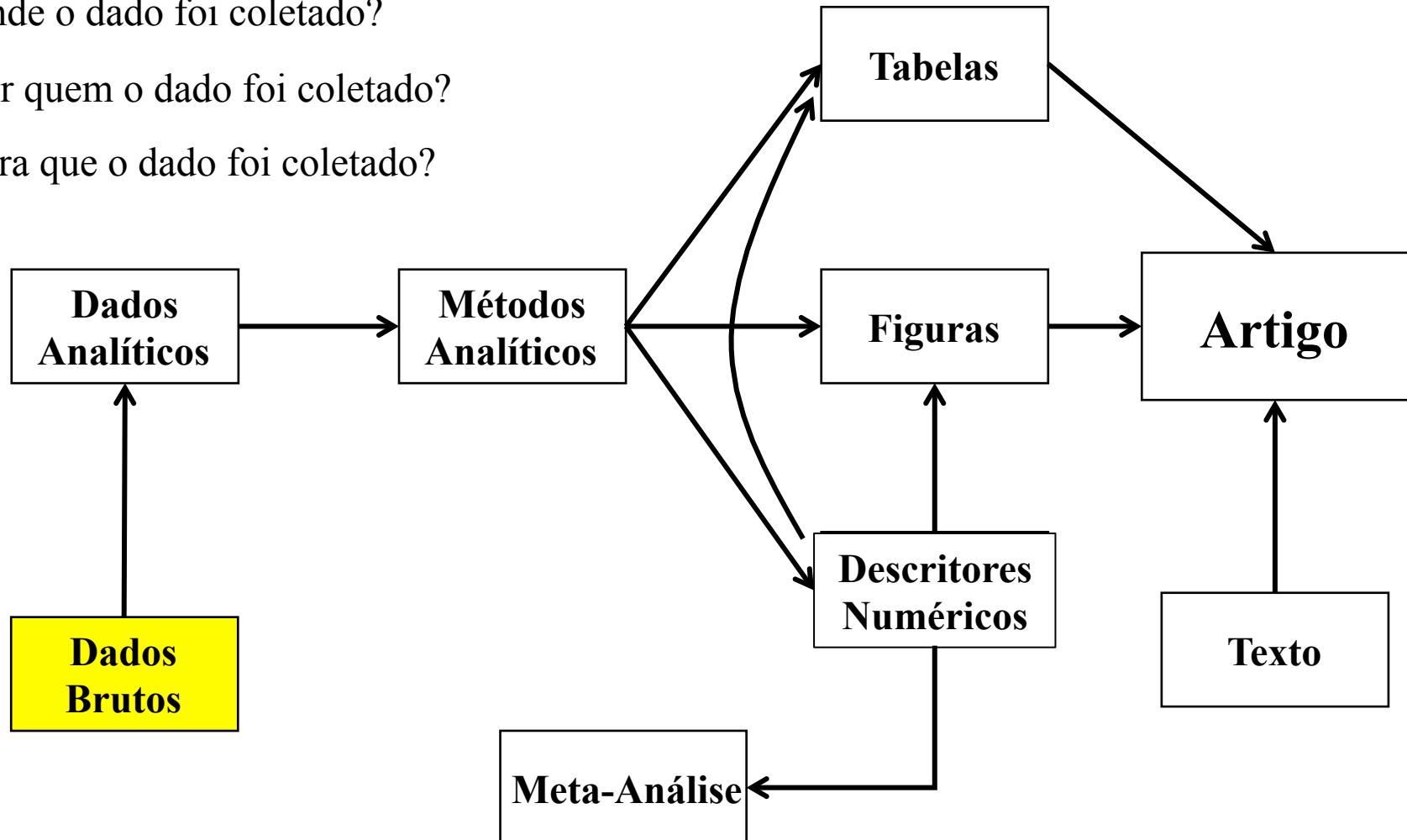
Uma nova Era para a ciência

- Qual a natureza dos dados?
- Controle de qualidade dos dados?
- Decisões sobre uso de um ou outro dado?



Uma nova Era para a ciência

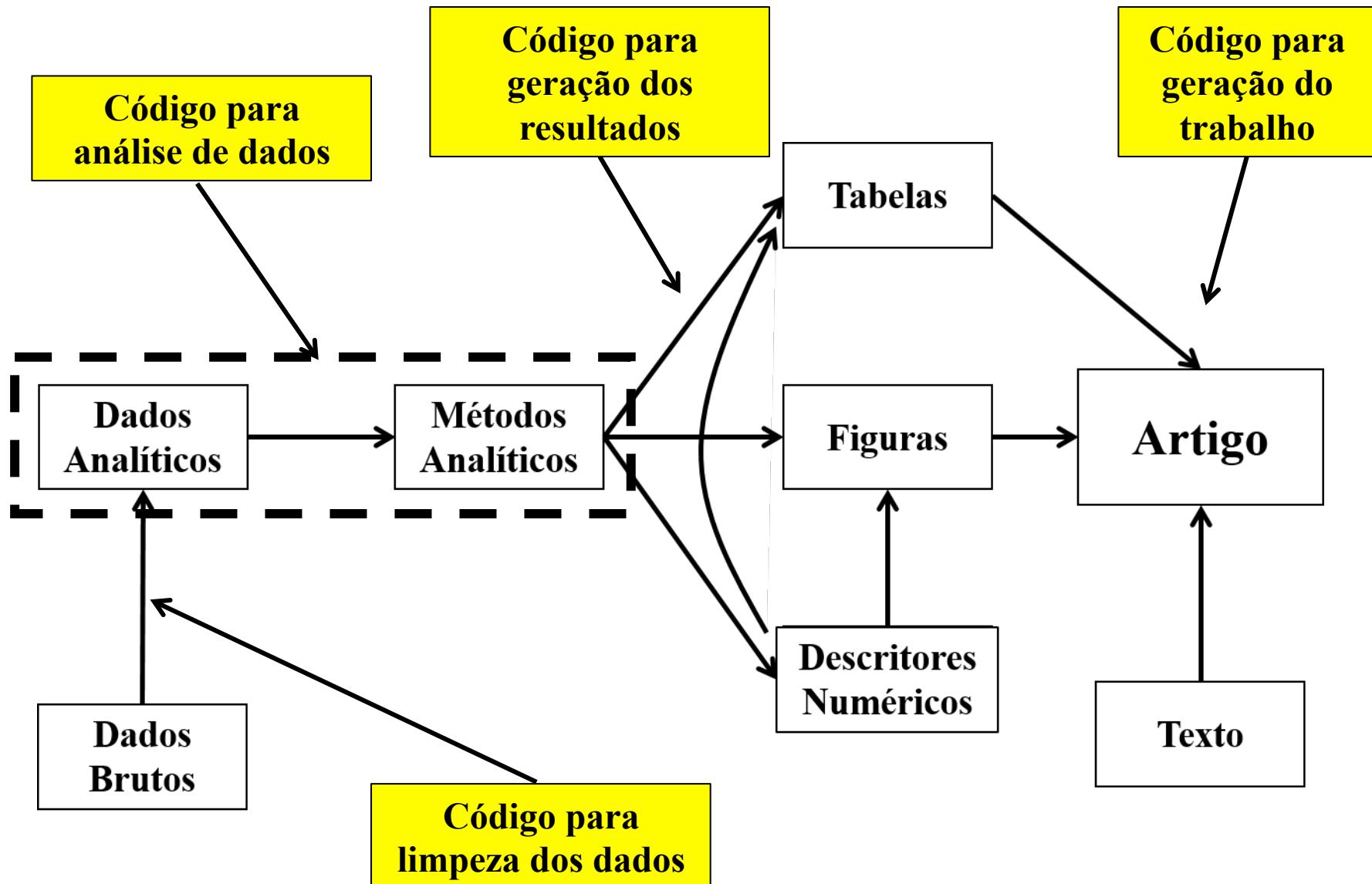
- Como o dado foi coletado?
- Onde o dado foi coletado?
- Por quem o dado foi coletado?
- Para que o dado foi coletado?



Tão importante quanto o artigo é o dado que o gerou

- *Open data*: compartilhamento e abertura de dados, pelo menos quando da publicação do trabalho.
 - ✓ Torna o processo de produção científica mais transparente;
 - ✓ Dá mais robustez aos seus achados;
 - ✓ Reduz a possibilidade de fraude na criação de dados;
 - ✓ Na maioria das vezes, *torna público o que é público*: a maior parte da pesquisa científica é financiada com dinheiro público;
 - ✓ A abertura da disponibilidade de dados pode ajudar na conservação de espécies e ecossistemas.

Não só os dados, mas também o “código” usado



Uma Era de *Open Science*

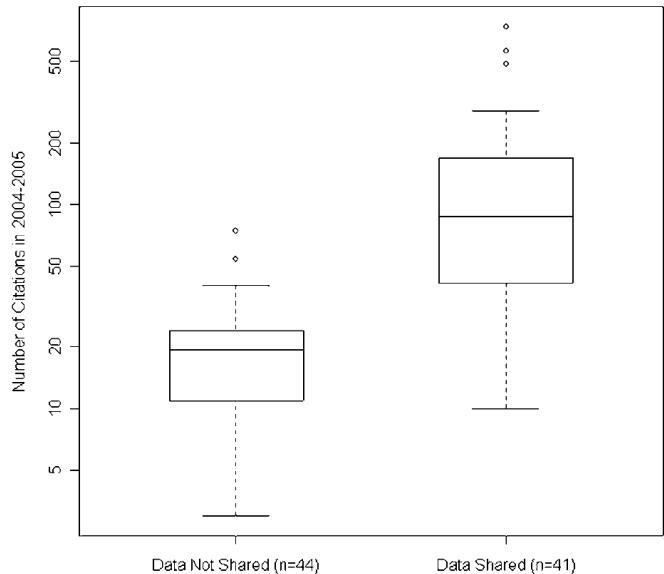
- O compartilhamento dos dados e do código utilizado é importante para garantir a reprodução do trabalho, agora e depois.
- A chance de um dado se perder aumenta com o tempo: reinventando a roda.

Table 1. Breakdown of Data Availability by Year of Publication

Year	No Working E-Mail	No Response to E-Mail	Response Did Not Give Status of Data	Data Lost	Data Exist, Unwilling to Share	Data Received
1991	9 (35%)	9 (35%)	2 (8%)	4 (15%)	1 (4%)	1 (4%)
1993	14 (39%)	11 (31%)	3 (8%)	7 (19%)	0 (0%)	1 (3%)
1995	11 (31%)	9 (26%)	0 (0%)	7 (20%)	2 (6%)	6 (17%)
1997	11 (37%)	9 (30%)	1 (3%)	2 (7%)	3 (10%)	4 (13%)
1999	19 (48%)	13 (32%)	1 (2%)	1 (2%)	0 (0%)	6 (15%)
2001	13 (30%)	15 (35%)	3 (7%)	4 (9%)	0 (0%)	8 (19%)
2003	9 (20%)	20 (43%)	4 (9%)	2 (4%)	0 (0%)	11 (24%)
2005	11 (24%)	14 (31%)	6 (13%)	1 (2%)	0 (0%)	13 (29%)
2007	12 (18%)	31 (47%)	2 (3%)	4 (6%)	1 (2%)	16 (24%)
2009	9 (13%)	34 (49%)	3 (4%)	5 (7%)	6 (9%)	12 (17%)
2011	13 (16%)	29 (36%)	8 (10%)	0 (0%)	7 (9%)	23 (29%)
Totals	131 (25%)	194 (38%)	33 (6%)	37 (7%)	20 (4%)	101 (19%)

Uma Era de *Open Science*

- O compartilhamento dos dados e do código utilizado é importante para garantir a reprodução do trabalho, agora e depois.
- A chance de um dado se perder aumenta com o tempo: reinventando a roda;
- Trabalhos que compartilham os dados são, normalmente, mais citados;

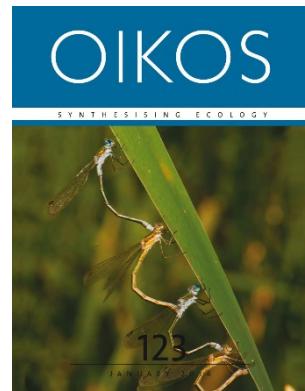
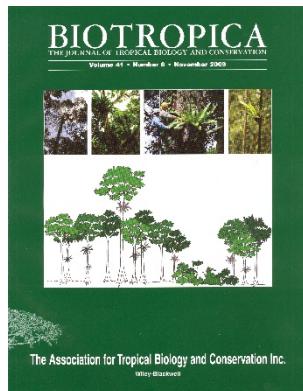
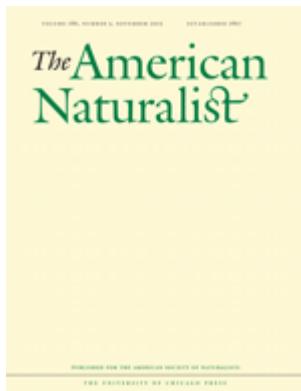


Você pode possivelmente não gostar, mas...

- Cientistas no início da carreira tendem a não querer compartilhar seus dados e/ou códigos (Tenopir et al, 2014, PLOS One):
 - ✓ “**P**osso ter cometido algum erro.” (mas se está errado, precisa ser consertado de qualquer forma, não?);
 - ✓ “**A**lguém pode publicar esses dados antes de mim” (não é ético, e a academia tem punições severas para pessoas que cometem esse deslize).
 - ✓ “**S**e alguém usar meu dado/código, não vai me dar crédito pela publicação” (você consegue dar um DOI para praticamente TUDO).

E, por mais que você não queira...

- Revistas na área de ecologia e evolução estão exigindo o compartilhamento de dados para a publicação dos artigos.



+ 99 revistas, cadastradas na
Dryad Digital Repository
(PLOS, Behavioral Ecology,
Ecography, Ecology, J Ecol, Funct
Ecol, Heredity, Evolution...)

- Existe a perspectiva de que, em breve, outras revistas vão aderir ao movimento – por exemplo, a Hydrobiologia.

E, por mais que você não queira...

- Além disso, revistas de outras áreas pedem o compartilhamento do código usado no trabalho e até “marcam” o seu artigo baseado na reproducibilidade das informações – Biostatistics.

Biostatistics (2014), 15, 1, pp. 1–12
doi:10.1093/biostatistics/kxt007
Advance Access publication on September 25, 2013

An estimate of the science-wise false discovery rate and application to the top medical literature

LEAH R. JAGER

Department of Mathematics, United States Naval Academy, Annapolis, MD 21402, USA

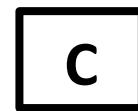
JEFFREY T. LEEK*

*Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205,
USA*

jleek@jhsph.edu



Dados são compartilhados



Código é compartilhado



Dados e código são compartilhados e o editor associado é capaz de reproduzir os resultados apresentados no artigo.

Onde compartilhar seus dados?

- Fora do Brasil, além das revistas, os órgãos de fomento pedem (1) o compartilhamento dos dados e (2) um plano para armazenamento e disponibilização destes dados em longo prazo.



zenodo



Uma pedra no meio do caminho

- Nós não somos treinados em uma cultura de compartilhamento de dados;

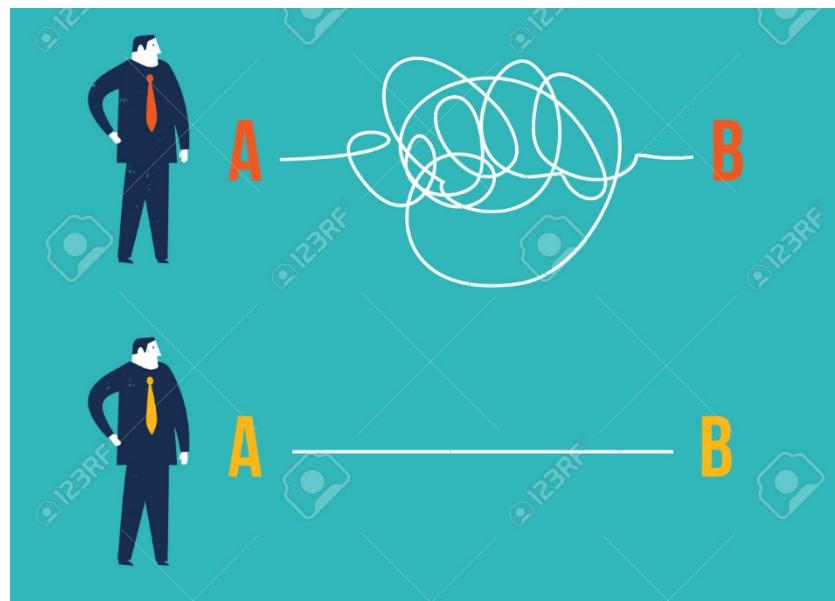
Algumas pedras no meio do caminho

- Nós não somos treinados em uma cultura de compartilhamento de dados;
- Nossa habilidade de organização é, normalmente, ruim;



Algumas pedras no meio do caminho

- Nós não somos treinados em uma cultura de compartilhamento de dados;
- Nossa habilidade de organização é, normalmente, ruim;
- Relatamos mal o que fazemos;



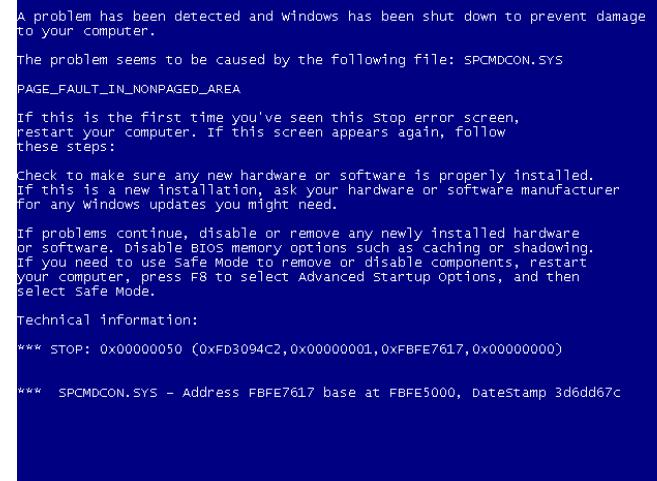
Algumas pedras no meio do caminho

- Nós não somos treinados em uma cultura de compartilhamento de dados;
- Nossa habilidade de organização é, normalmente, ruim;
- Relatamos mal o que fazemos;
- Gastamos muito tempo fazendo, pouco tempo pensando;



Algumas pedras no meio do caminho

- Nós não somos treinados em uma cultura de compartilhamento de dados;
- Nossa habilidade de organização é, normalmente, ruim;
- Relatamos mal o que fazemos;
- Gastamos muito tempo fazendo, pouco tempo pensando;
- Não temos tanta habilidade computacional.



Cuidando do seu trabalho

- Você pode usar ferramentas *online* para facilitar o seu trabalho, seja de forma textual ou através da criação de fluxogramas de trabalho:
 - ✓ **myExperiment**: <http://www.myexperiment.org/home>
 - ✓ **Kepler**: <https://kepler-project.org/>
 - ✓ **Taverna**: <http://www.taverna.org.uk/>

my experiment

What we do

myExperiment is a collaborative environment where scientists can safely publish their workflows and *in silico* experiments, share them with groups and find those of others. **Workflows**, other digital objects and bundles (called **Packs**) can now be swapped, sorted and searched like photos and videos on the Web. Unlike Facebook or MySpace, myExperiment fully understands the needs of the researcher and makes it really easy for the next generation of scientists to contribute to a pool of scientific methods, build communities and form relationships — reducing time-to-experiment, sharing expertise and avoiding reinvention. myExperiment is now the largest public repository of scientific workflows.

Using myExperiment

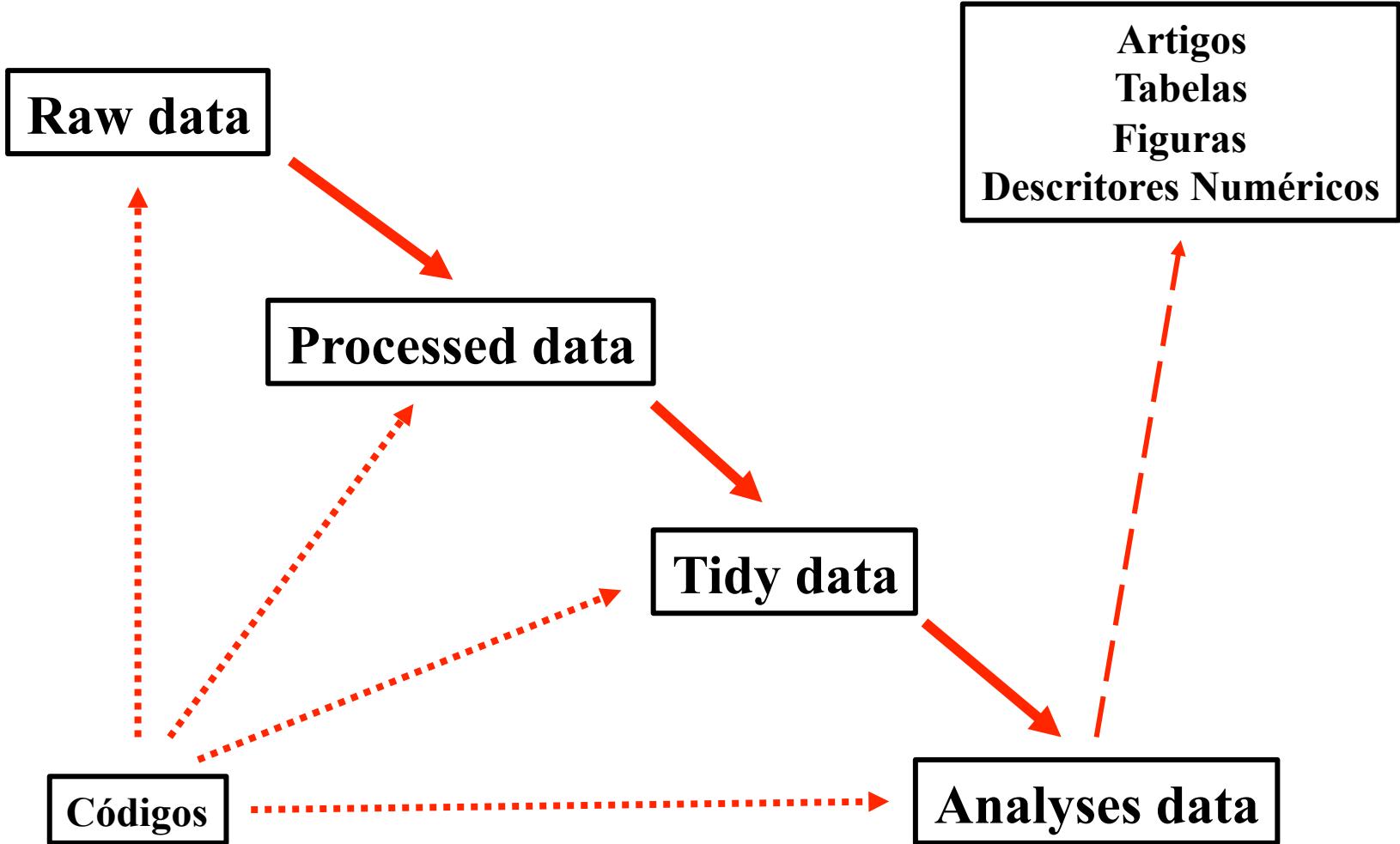
You can **find workflows** here; myExperiment has many different types of workflows, such as Taverna, Galaxy, Rapid Miner, Bio Extract, and Kepler. You can also **share your workflows** and their **supporting files**, either with your colleagues or with the wider world, or even **make packs** that group together related workflows and files, allowing people to download a whole body of work with a single click.

Of course, you can also **find the other people** on myExperiment, and **create and join groups** of people to collaborate. We also support **tagging** of workflows and files, and you can even **review others' contributions** right here on myExperiment.

myExperiment makes it easy to **find, use and share scientific workflows** and other **Research Objects**, and to build communities.



De volta ao início



*Dados analíticos = Processed + Tidy + Analyses data

Otimize o seu tempo

- **Don't repeat yourself** - automatize:
 - ✓ Processos repetitivos (fazer a mesma operação n vezes com dados diferentes);
 - ✓ Cada passo do seu trabalho (manipular uma planilha para gerar outra);
- **Use a tecnologia com sabedoria:** você pode perder horas para aprender algo que vai te poupar dias de trabalho no futuro *vs* algo que só será usado poucas vezes.

Seja prático com suas planilhas de dados

Se for adicionar informações à planilha no futuro, utilize linhas para isso, não colunas.

PIOR

variavel	site1	site2	site3
Sp1	0	5	8
Sp2	10	3	0
var1	80.7	30	10
var2	9	20	30

MELHOR

Site_ID	Sp1	Sp2	var1	var2
Site1	0	10	80.7	9
Site2	5	3	30	20
Site3	8	0	10	30

Seja prático com suas planilhas de dados

Cada planilha deve conter um tipo de variável, com uma coluna ‘chave’ entre elas.

Site_ID	sp	abundancia
Site1	1	0
Site2	1	5
Site3	1	8
Site1	2	10
Site2	2	3
Site3	2	0

Site_ID	variavel	valor
Site1	var1	80.7
Site2	var1	30
Site3	var1	10
Site1	var2	9
Site2	var2	20
Site3	var2	30

Seja prático com suas planilhas de dados

Use nomes mais intuitivos para as suas variáveis (depois de um tempo, isso faz a diferença).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	brom	brom_sp	diam	temp_max	chl	turb	vol_max	leafs	cor	copos	vol	ph	brom_nb	pos
2	VN1	VN		50	35	0.1672	47.56	1630	67	0.012	20	420	4.51	1 exp
3	VN2	VN		37	38	0.2568	33.36	1000	45	0.025	16	80	5.42	1 exp
4	VN3	VN		46	35	2.4813	333.6	2000	70	0.017	22	200	4.95	1 exp
5	VN4	VN		45	32	1.6218	615.9	1570	65	0.096	23	325	4.44	1 exp
6	VN5	VN		40	28	1.7139	64.27	1400	80	0.041	25	165	4.48	1 exp
7	VN6	VN		20	20	0.0522	40.40	710	42	0.026	16	115	4.74	1 exp

VS

	IDENTIFICAÇÃO			VARIÁVEL RESPOSTA				TR		
1	id_estudo	par_id	nivel_id	VAR_study	VAR_unity	VAR_ID	VAR_source	CG_stressor	CG_name_study	CG_treat
2	22	362	Controle	Resource consumed	no	197	Figure 2a	Warming	North	Controle
3	22	362	Tratamento	Resource consumed	no	197	Figure 2a	Warming	Central	Warming
4	22	363	Controle	Resource consumed	no	197	Figure 2a	Warming	North	Controle

Seja prático com suas planilhas de dados

Use nomes mais intuitivos para as suas planilhas de dados.

Name	Type	Size	Date modified
distsam	Text Document	2 KB	19/04/2011 22:13
distpcnm	Text Document	28 KB	19/04/2011 22:28
distfiltro	Text Document	4 KB	19/04/2011 22:19
distancia	Text Document	25 KB	19/04/2011 13:31
brom_fuzzy	Text Document	1 KB	09/05/2011 20:59
bioticos	Text Document	6 KB	20/04/2011 12:48
ambiental	Text Document	4 KB	13/06/2011 08:31

vs

Name	Type	Size	Date modified
abioticos	Planilha do Micro...	369 KB	19/04/2016 11:33
bentos	Planilha do Micro...	1,971 KB	26/03/2016 21:27
fitoplancton	Planilha do Micro...	412 KB	26/03/2016 21:29
peixes	Planilha do Micro...	378 KB	26/03/2016 21:33
species traits	Planilha do Micro...	72 KB	23/03/2016 16:39
zooplankton	Planilha do Micro...	215 KB	26/03/2016 21:31

Mantenha seus dados organizados

Use uma pasta para cada tipo de informação/projeto.

Name	Date modified	Type
Figuras	15/02/2015 23:57	File folder
Input	09/07/2015 07:34	File folder
Resultados	09/07/2015 07:34	File folder
Script	07/12/2015 19:48	File folder

Name	Date modified	Type
auxiliary files	19/04/2016 11:33	File folder
data for analyzes	23/03/2016 18:17	File folder
data output	29/03/2016 15:51	File folder
processed data	19/03/2016 16:13	File folder
R functions	25/03/2016 17:21	File folder
R Markdown	19/04/2016 11:34	File folder
R scripts to analyze	19/04/2016 11:34	File folder
R scripts to tidy	22/03/2016 19:05	File folder
raw data	20/04/2016 13:14	File folder
shapefiles	19/04/2016 11:34	File folder

Cuide do seu código

Antes de começar a escrever qualquer código, pense primeiro no que você quer fazer.

1. Carregar/simular dados;
2. Carregar pacotes que vou usar;
3. Criar um código para as categorias;
4. Transformar o formato da tabela;
5. Fazer um cálculo;
6. Desenhar um gráfico
7. Salvar o gráfico.

Cuide do seu código

Minimize a sua intervenção no código cada vez que você for fazer
alguma coisa.

NUNCA USE file.choose() PARA LER UM ARQUIVO NO R

```
fitoplancton <- read.table("processed data/fitoplancton.txt", header = TRUE)
zooplankton <- read.table("processed data/zooplankton.txt", header = TRUE)
bentos <- read.table("processed data/bentos.txt", header = TRUE)
abioticos <- read.table("processed data/abiotic.txt", header = TRUE)
species <- readxl::read_excel("raw data/species traits.xlsx")
```

Cuide do seu código

Evite nomear objetos:

- Com nomes de funções já existentes no ambiente R (data <-dados);
- Como nomes de argumentos já existentes nas funções;
- Com nomes muito complicados (planilhaAnalisedeDadosnova2a);
- Com nomes pouco informativos (m1 <- <resultado de um modelo>).

DICA: sempre aperte o ‘tab’ quando começar a escrever o nome que você quer dar para um objeto – você vai ter acesso a toda a lista de nomes parecidos com aqueles; objetos com nomes em português normalmente não dão problemas de conflito.

Cuide do seu código

Crie um projeto para todo o trabalho que você for fazer, usando R ou não: **o R entende que toda a informação necessária para um trabalho está contida no diretório onde o projeto foi criado, independente do computador em que esteja.**

Name	Date modified	Type
auxiliary files	19/04/2016 11:33	File folder
data for analyzes	23/03/2016 18:17	File folder
data output	29/03/2016 15:51	File folder
processed data	19/03/2016 16:13	File folder
R functions	25/03/2016 17:21	File folder
R Markdown	19/04/2016 11:34	File folder
R scripts to analyze	19/04/2016 11:34	File folder
R scripts to tidy	22/03/2016 19:05	File folder
raw data	20/04/2016 13:14	File folder
shapefiles	19/04/2016 11:34	File folder
Igarapes	07/03/2016 12:10	Text Document
	07/05/2015 15:17	R Project

```
fitoplancton <- read.table("processed data/fitoplankton.txt", header = TRUE)
zooplankton <- read.table("processed data/zooplankton.txt", header = TRUE)
bentos <- read.table("processed data/bentos.txt", header = TRUE)
abioticos <- read.table("processed data/abiotic.txt", header = TRUE)
species <- readxl::read_excel("raw data/species traits.xlsx")
```

The diagram shows two arrows originating from the R code block and pointing to the 'processed data' and 'raw data' folders in the file list. The first arrow points to the 'processed data' folder, and the second arrow points to the 'raw data' folder.

Cuide do seu código

Um script, um objetivo: limpar dados, gerar dados para análise, gerar e testar modelos, gerar figuras, fazer download de arquivos,...;

Nome	^	Data de Modificação	Tamanho	Tipo
► to analyze data		12 de abr de 2017 23:13	--	Pasta
► to get data		16 de nov de 2016 20:04	--	Pasta
► to get results		19 de abr de 2017 23:38	--	Pasta

Nome	^	Data de Modificação	Tamanho
01 - importar dados.R		17 de nov de 2016 14:16	4 KB
02 - fitoplanton.R		6 de abr de 2017 16:36	15 KB
03 - zooplanton.R		6 de abr de 2017 16:42	6 KB
04 - bentos.R		6 de abr de 2017 17:05	19 KB
05 - peixes.R		6 de abr de 2017 18:29	14 KB
06 - funcionais.R		6 de abr de 2017 17:44	8 KB
07 - indices.R		17 de nov de 2016 16:18	4 KB
08 - abioticos.R		6 de abr de 2017 17:49	4 KB

Cuide do seu código

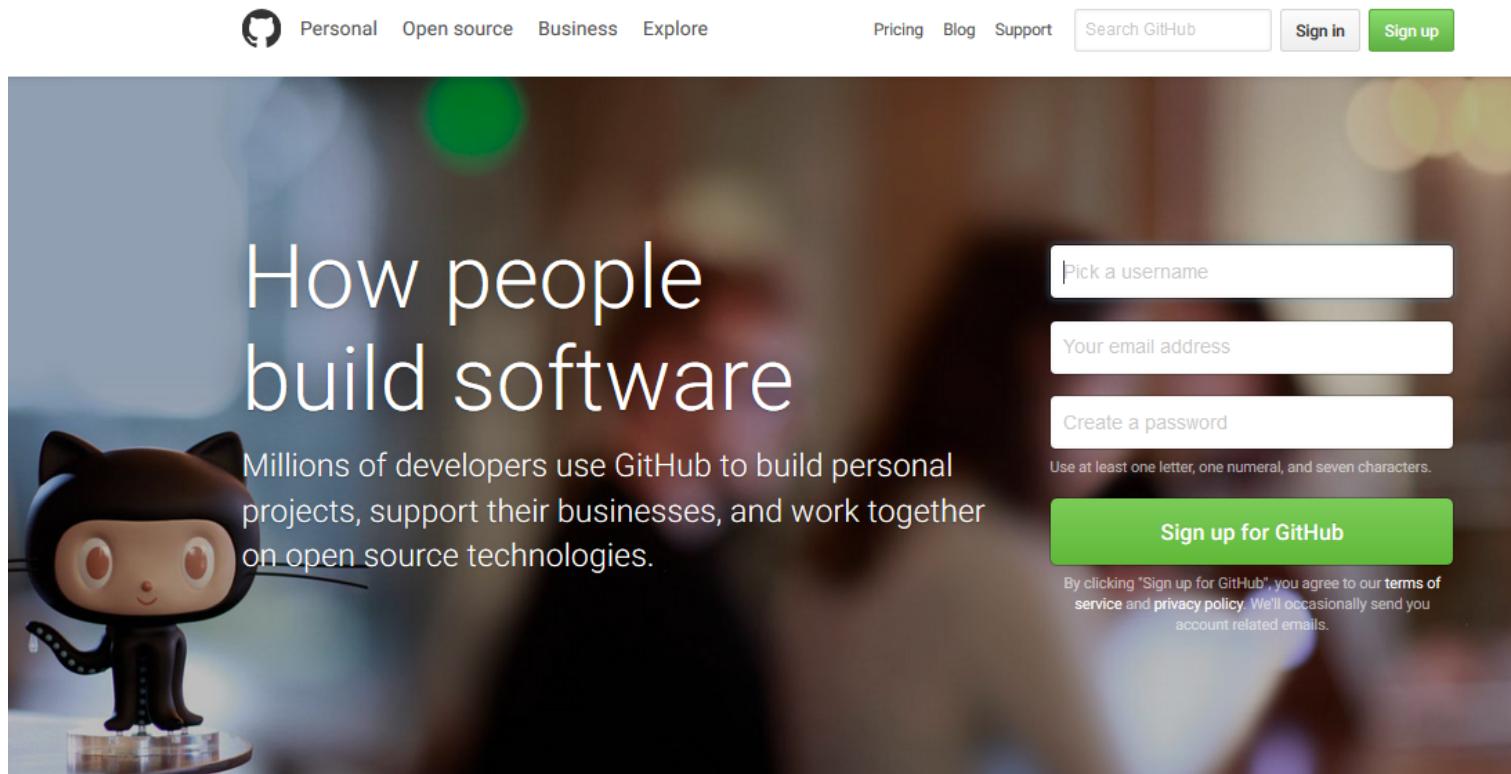
Quando for escrever qualquer código, lembre-se de:

- Usar o *indent*;
- Estabelecer “seções” no código;
- Comentar o que o código está fazendo.

```
# this function calculates the dissimilarity between pairs of sites, but taking into account differ
# in species richness in each comparison
# in order to improve the estimate of the rarefied dissimilarity, this rarefaction should be run fo
# number of times, from which we take the average dissimilarity between pairs of communities after
# .reps runs
beta.rich <- function(.dados, .reps){
  reps <- .reps
  #create an empty data frame where you are going to store each pairwise dissimilarity value for ea
  jac0 <- as.data.frame(matrix(0, ncol=nrow(.dados), nrow=nrow(.dados)), row.names = row.names(.dad
  #create an empty data frame where you are going to sum up the values for each pairwise comparisor
  jac1 <- as.data.frame(matrix(0, ncol=nrow(.dados), nrow=nrow(.dados)), row.names = row.names(.dad
  for(k in 1:reps) {
    for(j in 1:nrow(.dados)) {
      for(i in j:nrow(.dados)) {
        #for sites that match their species richness, there is no need of rarefaction in the comparis
        if (sum(.dados[i,]) == sum(.dados[j,])) {
          r1 <- as.numeric(vegdist(.dados[c(i,j),], method="jac"))
          jac0[i,j] <- r1
        } else if (sum(.dados[i,]) > sum(.dados[j,])) { #if one sites is richer than the other, app
```

Facilite a gestão do seu projeto

Utilize um software de controle de versão (Git ou GitHub) e/ou de base de dados.



Facilite a gestão do seu projeto

Crie um arquivo de **metadata** sobre o seu trabalho:

- Metadata: dados sobre os dados – *por que, para que, quem, quando, onde, como, o que?*
- Alguns pacotes: ‘*EML*’ e ‘*datapackage*’.
- Ou faça você mesmo!

Informação	Conteúdo
Trabalho (Teórico)	Motivação & Perguntas
Proprietário	Quem coletou, endereço, telefone, e-mail, website
Trabalho (prático)	Quando foi coletado, onde, que horas, por quem
Método	Como o trabalho foi feito
Medidas	Como as medidas foram feitas
Dados	O que tem na planilha de dados, quais unidades, quais medidas

Resumindo

- Tão importante quanto seus resultados é como você chegou neles;
- O compartilhamento de informações (dados e código) é fundamental para o progresso científico;
- Você vai precisar compartilhar os dados do seu trabalho;
- Existem padrões e recomendações para que seu dados possam ser compartilhados de forma adequada, garantindo o uso futuro por outras pessoas e por você mesmo.
- Você precisa treinar seus colaboradores nessa cultura.